Terry D. Oswalt
Editor-in-Chief

William C. Keel
Volume Editor

Planets, Stars and Stellar Systems

VOLUME 6

Extragalactic Astronomy and Cosmology

Springer Reference

Planets, Stars and Stellar Systems

Extragalactic Astronomy and Cosmology

Terry D. Oswalt (Editor-in-Chief)

William C. Keel (Volume Editor)

Planets, Stars and Stellar Systems

Volume 6:
Extragalactic Astronomy and Cosmology

With 314 Figures and 12 Tables

Springer Reference

Editor-in-Chief
Terry D. Oswalt
Department of Physics & Space Sciences
Florida Institute of Technology
University Boulevard
Melbourne, FL, USA

Volume Editor
William C. Keel
Department of Physics and Astronomy
University of Alabama
Tuscaloosa, AL, USA

ISBN 978-94-007-5608-3 ISBN 978-94-007-5609-0 (eBook)
ISBN 978-94-007-5610-6 (print and electronic bundle)
DOI 10.1007/978-94-007-5609-0

This title is part of a set with
Set ISBN 978-90-481-8817-8
Set ISBN 978-90-481-8818-5 (eBook)
Set ISBN 978-90-481-8852-9 (print and electronic bundle)

Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012953926

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Series Preface

It is my great pleasure to introduce "Planets, Stars, and Stellar Systems" (PSSS). As a "Springer Reference", PSSS is intended for graduate students to professionals in astronomy, astrophysics and planetary science, but it will also be useful to scientists in other fields whose research interests overlap with astronomy. Our aim is to capture the spirit of 21st century astronomy – an empirical physical science whose almost explosive progress is enabled by new instrumentation, observational discoveries, guided by theory and simulation.

Each volume, edited by internationally recognized expert(s), introduces the reader to a well-defined area within astronomy and can be used as a text or recommended reading for an advanced undergraduate or postgraduate course. Volume 1, edited by Ian McLean, is an essential primer on the tools of an astronomer, i.e., the telescopes, instrumentation and detectors used to query the entire electromagnetic spectrum. Volume 2, edited by Howard Bond, is a compendium of the techniques and analysis methods that enable the interpretation of data collected with these tools. Volume 3, co-edited by Linda French and Paul Kalas, provides a crash course in the rapidly converging fields of stellar, solar system and extrasolar planetary science. Volume 4, edited by Martin Barstow, is one of the most complete references on stellar structure and evolution available today. Volume 5, edited by Gerard Gilmore, bridges the gap between our understanding of stellar systems and populations seen in great detail within the Galaxy and those seen in distant galaxies. Volume 6, edited by Bill Keel, nicely captures our current understanding of the origin and evolution of local galaxies to the large scale structure of the universe.

The chapters have been written by practicing professionals within the appropriate subdisciplines. Available in both traditional paper and electronic form, they include extensive bibliographic and hyperlink references to the current literature that will help readers to acquire a solid historical and technical foundation in that area. Each can also serve as a valuable reference for a course or refresher for practicing professional astronomers. Those familiar with the "Stars and Stellar Systems" series from several decades ago will recognize some of the inspiration for the approach we have taken.

Very many people have contributed to this project. I would like to thank Harry Blom and Sonja Guerts (Sonja Japenga at the time) of Springer, who originally encouraged me to pursue this project several years ago. Special thanks to our outstanding Springer editors Ramon Khanna (Astronomy) and Lydia Mueller (Major Reference Works) and their hard-working editorial team Jennifer Carlson, Elizabeth Ferrell, Jutta Jaeger-Hamers, Julia Koerting, and Tamara Schineller. Their continuous enthusiasm, friendly prodding and unwavering support made this series possible. Needless to say (but I'm saying it anyway), it was not an easy task shepherding a project this big through to completion!

Most of all, it has been a privilege to work with each of the volume Editors listed above and over 100 contributing authors on this project. I've learned a lot of astronomy from them, and I hope you will, too!

Terry D. Oswalt

January 2013

Terry D. Oswalt
General Editor

Preface to Volume 6

Intentionally, this compendium invites contrast with the similarly titled volume from 1975, edited by Alan and Mary Sandage and Jerome Kristian. They describe its view as largely a product of the late 1960s. Our understanding of galaxies then, despite the already enormous observational and theoretical effort summarized, now seems woefully incomplete, largely because we stand on the shoulders of technological innovators. We were then at the very dawn of X-ray observations of galaxies and clusters, and likewise our knowledge of the radio structures of galaxies and AGN was poised for dramatic improvements in quality and quantity. AGN were still a novelty whose connection to "ordinary" galaxies was almost unknown; the model of energy release during accretion into massive black holes cold not yet be clearly formulated.

As far as we knew then galaxy masses were baryonic, and there was something suspicious about virial mass estimates for galaxy clusters. Photometry of galaxies remained a tedious project, either using photomultipliers or the black art of calibrated photographic photometry. Digital instrumentation for optical astronomy was just beginning to appear; the basic techniques of observation had advanced only incrementally for many years. One catalog volume could still contain essentially all the photometric and redshift data ever obtained for galaxies. We knew almost nothing about the evolution of galaxies; the relevant observations remained closely mixed with the quest for the fundamental parameters of cosmology.

Since that volume appeared (during my undergraduate years), our view of galaxies and their context has broadened dramatically, sometimes in ways scarcely conceivable then.

It seems appropriate to contrast many of these views to our current picture. Not only can we trace the history of star formation (galaxy evolution seen in the act) across cosmic time, but we can address this question with constraints all the way from the X-ray to radio regimes, coupling direct detection of young stellar populations with the secondhand emission from dust grains and supernova products.

We continue to find value in the galaxy categories bequeathed by Edwin Hubble, Alan Sandage, and Sidney van den Bergh, but now extend the study of galaxy structure across cosmic time and wavelength. Connections emerge (some hotly debated) between details of galaxy structure and events in galaxy history. Again and again, crucial observable properties of galaxies are seen to be driven by dark matter, whose properties are being narrowed by such techniques as gravitational lensing.

With the finding of supermassive black holes (and thus potentially "dead quasars") in most luminous galaxies, we seem to have the answer to a question posed by Joe Miller decades ago-are quasars important or merely interesting? If galaxies are important, so are quasars.

Our understanding of clusters and larger-scale structures has been revolutionized, both with the finding that the hot intracluster medium carries more mass than the stars, and with broad surveys showing statistical properties of superclusters, voids, and filaments (it is gratifying to note that one figure in this volume shows a reprocessing of the same galaxy counts shown in contour form by C.D. Shane in 1975). Further, we are starting to fill in physical detail as to the ways galaxies are affected by their environments. At the level of detail we can now reach, no galaxy is really an island Universe.

The cosmic distance scale remains important, but we are in a very different stage. A web of interlocking as well as independent measurements has narrowed the value of the local Hubble constant to a few per cent, so that local motions superimposed on the expansion are measurable.

We are at a key juncture in cosmology. Recent results have dramatically narrowed the values of the Hubble constant, mass density, and fluctuation amplitude in the early Universe, with further refinements expected soon from the Planck mission. Fine structure in the microwave background radiation encodes a rich range of both physical and astrophysical processes. Yet we still only anticipate a physical understanding of whatever causes the acceleration of cosmic expansion indicated by multiple techniques. Such terms as dark energy, cosmological constant, and quintessence at this point serve mostly to organize our ignorance.

Likewise, with new facilities capable of narrowing the observational Dark Ages, we see promise of adding new bodies of data to underpin our understanding of galaxy formation, and the first stars, and the growth of black holes in the early Universe.

All readers of this collection owe a debt to the community spirit of the authors, who have invested so much time and effort into making their contributions. I hope that this collection shares with its predecessor a long useful life for many chapters, but also the scientific joy of being overtaken in some aspects by utterly unexpected discoveries.

William C Keel

William C. Keel
Tuscaloosa, Alabama
USA

Editor-in-Chief

Dr. Terry D. Oswalt
Department Physics & Space Sciences
Florida Institute of Technology
150 W. University Boulevard
Melbourne, Florida 32901
USA
E-mail: toswalt@fit.edu

Dr. Oswalt has been a member of the Florida Tech faculty since 1982 and was the first professional astronomer in the Department of Physics and Space Sciences. He serves on a number of professional society and advisory committees each year. From 1998 to 2000, Dr. Oswalt served as Program Director for Stellar Astronomy and Astrophysics at the National Science Foundation. After returning to Florida Tech in 2000, he served as Associate Dean for Research for the College of Science (2000–2005) and interim Vice Provost for Research (2005–2006). He is now Head of the Department of Physics & Space Sciences. Dr. Oswalt has written over 200 scientific articles and has edited three astronomy books, in addition to serving as Editor-in-Chief for the six-volume Planets, Stars, and Stellar Systems series.

Dr. Oswalt is the founding chairman of the Southeast Association for Research in Astronomy (SARA), a consortium of ten southeastern universities that operates automated 1-meter class telescopes at Kitt Peak National Observatory in Arizona and Cerro Tololo Interamerican Observatory in Chile (see the website www.saraobservatory.org for details). These facilities, which are remotely accessible on the Internet, are used for a variety of research projects by faculty and students. They also support the SARA Research Experiences for Undergraduates (REU) program, which brings students from all over the U.S. each summer to participate one-on-one with SARA faculty mentors in astronomical research projects. In addition, Dr. Oswalt secured funding for the 0.8-meter Ortega telescope on the Florida Tech campus. It is the largest research telescope in the State of Florida.

Dr. Oswalt's primary research focuses on spectroscopic and photometric investigations of very wide binaries that contain known or suspected white dwarf stars. These pairs of stars, whose separations are so large that orbital motion is undetectable, provide a unique opportunity to explore the low luminosity ends of both the white dwarf cooling track and the main sequence; to test competing models of white dwarf spectral evolution; to determine the space motions, masses, and luminosities for the largest single sample of white dwarfs known; and to set a lower limit to the age and dark matter content of the Galactic disk.

Volume Editor

Dr. William C. Keel
Department of Physics and Astronomy
University of Alabama
Box 870324
206 Gallalee Hall
Tuscaloosa, AL 35487
USA

William C. Keel is Professor of Astronomy at the University of Alabama in Tuscaloosa. His astronomical interests began as a youngster using a secondhand reflector in the back yard, and he remains active as an amateur as well as professional astronomer. His undergraduate work was at Vanderbilt University, followed by a Ph.D. from the University of California, Santa Cruz. Dr. Keel had postdoctoral positions at Kitt Peak National Observatory and at Leiden, before taking up a faculty position in Alabama. His research interests span the galaxies - active galactic nuclei, galaxy interactions and evolution, dust in galaxies. Observationally oriented, his work has used spectral bands from the radio to the X-ray regimes, with the strongest emphasis in the optical and ultraviolet. These results have been reported in 150 refereed papers.

In recent years, much of his work has been tied to the enormously successful Galaxy Zoo citizen-science project. He continues to have scheduling responsibilities for the two telescopes of the SARA consortium, and has served on numerous NASA review panels including two Senior Reviews.

Dr. Keel has been active in outreach beyond the formal classroom, through magazine articles, an online presence in several discussion forums, and webcomics explaining ongoing Hubble Space Telescope programs.

He has written a technical monograph, The Road to Galaxy Formation, and the nontechnical volume The Sky at Einstein's Feet tracing the impact of relativity throughout astronomy.

Table of Contents

Series Preface v

Preface to Volume 6 vii

Editor-in-Chief ix

Volume Editor xi

List of Contributors xv

Volume 6

1 **Galaxy Morphology** 1
Ronald J. Buta

2 **Elliptical and Disk Galaxy Structure and Modern Scaling Laws** 91
Alister W. Graham

3 **Star Formation in Galaxies** 141
Samuel Boissier

4 **The Cool ISM in Galaxies** 183
Jan M. van der Hulst · W. J. G. de Blok

5 **The Influence of Environment on Galaxy Evolution** 207
Bernd Vollmer

6 **Clusters of Galaxies** 265
Richard Bower

7 **Active Galactic Nuclei** 305
Eric S. Perlman

8 **The Large-Scale Structure of the Universe** 387
Alison L. Coil

9 **The Distance Scale of the Universe** 423
Wendy L. Freedman · Barry F. Madore

10 **Galaxies in the Cosmological Context** 451
Gabriella De Lucia

11 Evolution of Active Galactic Nuclei 503
Andrea Merloni · Sebastian Heinz

12 The Intergalactic Medium 567
Renyue Cen

13 Cosmic Microwave Background... 609
John Mather · Gary Hinshaw · Lyman Page

Index 685

List of Contributors

W. J. G. de Blok
Astronomy Group
ASTRON
Netherlands Foundation for Radio
Astronomy
Dwingeloo
The Netherlands

Samuel Boissier
Laboratoire d'Astrophysique de Marseille
Université Aix-Marseille & CNRS
UMR7326
Marseille cedex 13
France

Richard Bower
Department of Physics
University of Durham
Durham
UK

Ronald J. Buta
Department of Physics and Astronomy
University of Alabama
Tuscaloosa, AL
USA

Renyue Cen
Department of Astrophysical Sciences
Princeton University Observatory
Peyton Hall
Princeton, NJ
USA

Alison L. Coil
Department of Physics
University of California
San Diego, CA
USA

Wendy L. Freedman
Carnegie Observatories
Pasadena
CA
USA

Alister W. Graham
Centre for Astrophysics and
Supercomputing
Swinburne University of Technology
Hawthorn
Australia

Sebastian Heinz
Astronomy Department
University of Wisconsin-Madison
Madison, WI
USA

Gary Hinshaw
Department of Physics and Astronomy
University of British Columbia
Vancouver, BC
Canada

Jan M. van der Hulst
Radio Astronomy
Kapteyn Astronomical Institute
University of Groningen
Groningen
The Netherlands

Gabriella De Lucia
INAF – Astronomical Observatory of Trieste
Trieste
Italy

Barry F. Madore
Carnegie Observatories
Pasadena, CA
USA

John Mather
Astrophysics Science Division
NASA/GSFC Code 443
Observational Cosmology
Greenbelt, MD
USA

Andrea Merloni
Max-Planck-Institut für Extraterrestrische Physik
Garching
Germany

Lyman Page
Department of Physics
Princeton University
Princeton, NJ
USA

Eric S. Perlman
Department of Physics and Space Sciences
Florida Institute of Technology
Melbourne, FL
USA

Bernd Vollmer
Observatoire astronomique de Strasbourg
Strasbourg
France

1 Galaxy Morphology

Ronald J. Buta
Department of Physics and Astronomy, University of Alabama,
Tuscaloosa, AL, USA

1 Introduction . . . 3

2 Overview . . . 4

3 Galaxy Classification . . . 6

4 A Continuum of Galactic Forms . . . 8

5 Galaxy Types: Stage, Family, and Variety . . . 10
5.1 Elliptical and Spheroidal Galaxies . . . 10
5.2 S0 and Spiral Galaxies . . . 14
5.3 Irregular Galaxies . . . 20

6 Other Dimensions to Galaxy Morphology . . . 21
6.1 Outer Rings and Pseudorings . . . 21
6.2 Inner and Outer Lenses . . . 24
6.3 Nuclear Rings and Bars . . . 24
6.4 Spiral Arm Morphologies . . . 27
6.5 Luminosity Effects . . . 30

7 The Morphology of Galactic Bars and Ovals . . . 31

8 Dust Morphologies . . . 35

9 The Morphologies of Galactic Bulges . . . 36

10 Effects of Interactions and Mergers . . . 39
10.1 Normal Versus Catastrophic Rings . . . 39
10.2 Environmental Effects on Star-Forming Disks . . . 42
10.3 Interacting and Peculiar Galaxies . . . 44
10.3.1 Tidal Tails, Arms, Bridges, and Streams . . . 44
10.3.2 Dust-Lane Ellipticals . . . 46
10.3.3 Shell/Ripple Galaxies . . . 46
10.3.4 Ultraluminous Infrared Galaxies . . . 47
10.4 Warps . . . 48
10.5 The Morphology of Active Galaxies . . . 49
10.6 The Morphology of Brightest Cluster Members . . . 52

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_1, © Springer Science+Business Media Dordrecht 2013

11 Star Formation Morphologies *53*
11.1 Hα Imaging 53
11.2 Ultraviolet Imaging 54
11.3 Atomic and Molecular Gas Morphology 55

12 Infrared Observations: Galactic Stellar Mass Morphology *59*

13 Intermediate and High Redshift Galaxy Morphology *65*

14 Giant Low-Surface-Brightness Galaxies *69*

15 Galaxy Morphology in Color *71*
15.1 Normal Galaxies 71
15.2 Dwarf Galaxies 72
15.2.1 dE, dS0, BCD, and cE Galaxies 73
15.2.2 Local Group Dwarf Spheroidals and Irregulars 74
15.2.3 Dwarf Spirals 75
15.3 Galaxy Zoo Project 76
15.4 Isolated Galaxies 77
15.5 Deep Field Color Imaging 79

16 Large-Scale Automated Galaxy Classification *79*

17 The Status and Future of Morphological Studies *81*

References *84*

Abstract: Hidden in the bewildering details of galaxy morphology are clues to how galaxies formed and have evolved over a Hubble time. This article reviews the phenomenology of galaxy morphology and classification using an extensive set of illustrations to delineate as many types as possible and to show how different types connect to various physical processes and characteristics. The old classification systems are refined, and new types introduced, as the explosion in available morphological data has modified our views on the structure and evolution of galaxies.

Keywords: Galaxies: Active, Galaxies: Classification, Galaxies: Clusters, Galaxies: Dwarfs, Galaxies: Elliptical, Galaxies: Galaxy Zoo project, Galaxies: High redshift, Galaxies: Isolated, Galaxies: Peculiar, Galaxies: S0s, Galaxies: Spiral, Galaxies: Structure

1 Introduction

In the nearly 100 years since galaxy morphology became a topic of research, much has been learned about galactic structure and dynamics. Known only as "nebulae" a century ago, galaxies were found to have a wide range of largely inexplicable forms whose relations to one another were a mystery. As data accumulated, it became clear that galaxies are fundamental units of matter in space, and an understanding of how they formed and evolved became one of the major goals of extragalactic studies. Even in the era of space observations, galaxy morphology continues to be the backbone of extragalactic research as modern instruments provide information on galactic structure across a wide range of distances and look-back times.

In spite of the advances in instrumentation and the explosion of data, classical galaxy morphology (i.e., the visual morphological classification in the style of Hubble and others) has not lost its relevance. The reasons for this are as follows:

1. Morphology is still a logical starting point for understanding galaxies. Sorting galaxies into their morphological categories is similar to sorting stars into spectral types and can lead to important astrophysical insights. Any theory of galaxy formation and evolution will have to, at some point, account for the vast array of galactic forms.
2. Galaxy morphology is strongly correlated with galactic star formation history. Galaxies where star formation ceased gigayears ago tend to look very different from those where star formation continues at the present time. Classical morphology recognizes these differences in an ordered way.
3. Information on galaxy morphology, in the form of new types of galaxies, multiwavelength views of previously known galaxy types, and higher resolution views of all or part of some galaxies, has exploded as modern instrumentation has superceded the old photographic plates that were once used exclusively for galaxy classification.
4. Galaxy classification has gone beyond the realm of a few thousand galaxies to that of a *million* galaxies through the Galaxy Zoo project. Not only this, but Galaxy Zoo has taken morphology from the exclusive practice of a few experts to the public at large, thus facilitating citizen science at its best. Galaxy Zoo images are also in *color*, thus allowing the recognition of special galaxy types and features based on stellar populations or gaseous emission.
5. Finally, deep surveys with the Hubble Space Telescope have extended morphological studies well beyond the realm of the nearby galaxies that dominated early catalogues, allowing detailed morphology to be distinguished at unprecedented redshifts.

Now, more than ever, galaxy morphology is a vibrant subject that continues to provide surprises as more galaxies are studied for their morphological characteristics across the electromagnetic spectrum. It is clear that a variety of effects are behind observed morphologies, including initial protogalactic cloud conditions, environmental density and merger/interaction history, internal perturbations, gas accretions, nuclear activity, properties of the dark matter halo, secular evolution, as well as the diversity in star formation histories, and that a global perspective based on large numbers of galaxies will improve theoretical models and give a more reliable picture of galactic evolution.

The goal of this chapter is to present the phenomenology of galaxy morphology in an organized way and highlight recent advances in understanding what factors influence morphology and how various galaxy types are interpreted. This chapter is a natural follow-up to the excellent review of galaxy morphology and classification by Sandage (1975) in Volume IX of the classic *Stars and Stellar Systems* series. It also complements the recently published *de Vaucouleurs Atlas of Galaxies* (Buta et al. 2007, hereafter the dVA), which provided a detailed review of the state and technique of galaxy classification up to about the year 2005. Illustrations are very important in a review of this nature, and this chapter draws on a large number of sources of images. For this purpose, the Sloan Digital Sky Survey (SDSS), the NASA/IPAC Extragalactic Database (NED), and the dVA have been most useful.

2 Overview

As extended objects rather than point sources, galaxies show a wide variety of forms, some due to intrinsic structures, others due to the way the galaxy is oriented to the line of sight. The random orientations and the wide spread of distances are the principal factors that can complicate interpretations of galaxy morphology. If we could view every galaxy along its principal axis of rotation, and from the same distance, then fairer comparisons would be possible. Nevertheless, morphologies seen in face-on galaxies can also often be recognized in more inclined galaxies (➔ *Fig. 1-1*). It is only for the highest inclinations that morphology switches from face-on radial structure to vertical structure. In general, we either know the planar structure in a galaxy or we know its vertical structure, but we usually cannot know both well from analysis of images alone.

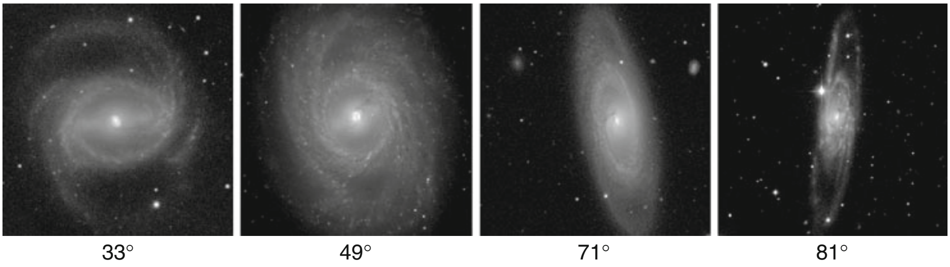


Fig. 1-1
Four galaxies of likely similar face-on morphology viewed at different inclinations (number below each image). The galaxies are (*left* to *right*): NGC 1433, NGC 3351, NGC 4274, and NGC 5792. Images are from the dVA (filters *B* and *g*)

Galaxy morphology began to get interesting when the "Leviathan of Parsonstown," the 72-in. meridian-based telescope built in the 1840s by William Parsons, Third Earl of Rosse, on the grounds of Birr Castle in Ireland, revealed spiral patterns in many of the brighter Herschel and Messier "nebulae." The nature of these nebulae as galaxies was not fully known at the time, but the general suspicion was that they were star systems ("island universes") like the Milky Way, only too distant to be easily resolved into their individual stars. In fact, one of Parsons' motivations for building the "Leviathan" was to try and resolve the nebulae to prove this idea. The telescope did not convincingly do this, but the discovery of spiral structure itself was very important because such structure added to the mystique of the nebulae. The spiral form was not a random pattern and had to be significant in what it meant. The telescope was not capable of photography, and observers were only able to render what they saw with it in the form of sketches. The most famous sketch, that of M51 and its companion NGC 5195, has been widely reproduced in introductory astronomy textbooks.

While visual observations could reveal some important aspects of galaxy morphology, early galaxy classification was based on photographic plates taken in the blue region of the spectrum. Silver bromide dry emulsion plates were the staple of astronomy beginning in the 1870s and were relatively more sensitive to blue light than to red light. Later, photographs taken with Kodak 103a-O and IIa-O plates became the standard for galaxy classification. In this part of the spectrum, massive star clusters, dominated by spectral class O and B stars, are prominent and often seen to line the spiral arms of galaxies. These clusters can give blue-light photographs a great deal of detailed structure for classification. It is these types of photographs which led to the galaxy classification systems in use today.

In such photographs, we see many galaxies as a mix of structures. Inclined galaxies reveal the ubiquitous *disk* shape, the most highly flattened component of any galaxy. Studies of Doppler wavelength shifts in the spectra of disk objects (like HII regions and integrated star light) reveal that disks rotate differentially. If a galaxy is spiral, the disk is usually where the arms are found, and also where the bulk of interstellar matter is found. The radial luminosity profile of a disk is usually *exponential*, with departures from an exponential being due to the presence of other structures.

In the central area of a disk-shaped galaxy, there is also often a bright and sometimes less flattened mass concentration in the form of a *bulge*. The nature of bulges and how they form have been a topic of much recent research and are discussed further in ➋ Sect. 9. Disk galaxies range from virtually bulgeless to bulge-dominated. In the center, there may also be a conspicuous *nucleus*, a bright central concentration that was usually lost to overexposure in photographs. Nuclei may be dominated by ordinary star light, or may be *active*, meaning their spectra show evidence of violent gas motions.

Bars are the most important internal perturbations seen in disk-shaped galaxies. A bar is an elongated mass often made of old stars crossing the center. If spiral structure is present, the arms usually begin near the ends of the bar. Although most easily recognized in the face-on view, bars have generated great interest recently in the unique ways they can also be detected in the edge-on view. Not all bars are made exclusively of old stars. In some bulgeless galaxies, the bar has considerable gas and recent star formation.

Related to bars are elongated disk features known as *ovals*. Ovals usually differ from bars in lacking higher order Fourier components (i.e., have azimuthal intensity distributions that vary mainly as 2θ), but nevertheless can be major perturbations in a galactic disk. The entire disk of a galaxy may be oval, or a part of it may be oval. Oval disks are most easily detected if there is considerable light or structure at larger radii.

Rings are prominent features in some galaxies. Often defined by recent star formation, rings may be completely closed features or may be partial or open, the latter called "pseudorings." Rings can be narrow and sharp or broad and diffuse. It is particularly interesting that several kinds of rings are seen and that some galaxies can have as many as four recognizable ring features. *Nuclear rings* are the smallest rings and are typically seen in the centers of barred galaxies. *Inner rings* are intermediate-scale features that often envelop the bar in a barred galaxy. Outer rings are large, low-surface-brightness features that typically lie at about twice the radius of a bar. Other kinds of rings, called accretion rings, polar rings, and collisional rings, are also known but are much rarer than the inner, outer, and nuclear rings of barred galaxies. The latter kinds of rings are also not exclusive to barred galaxies but may be found also in nonbarred galaxies.

Lenses are features, made usually of old stars, that have a shallow brightness gradient interior to a sharp edge. They are commonly seen in Hubble's disk-shaped S0 class (➜ Sect. 5.2). If a bar is present, the bar may fill a lens in one dimension. Lenses may be round or slightly elliptical in shape. If elliptical in shape, they would also be considered ovals.

Nuclear bars are the small bars occasionally seen in the centers of barred galaxies, often lying within a nuclear ring. When present in a barred galaxy, the main bar is called the "primary bar" and the nuclear bar is called the "secondary bar." It is possible for a nuclear bar to exist in the absence of a primary bar.

Dust lanes are often seen in optical images of spiral galaxies and may appear extremely regular and organized. They are most readily evident in edge-on or highly inclined disk galaxies but are still detectable in the face-on view, often on the leading edges of bars or the concave sides of strong inner spiral arms.

Spiral arms may also show considerable morphological variation. Spirals may be regular 1-, 2-, 3-, or 4-armed patterns and may also be higher order multiarmed patterns. Spirals may be tightly wrapped (low pitch angle) or very open (high pitch angle). A *grand-design* spiral is a well-defined global pattern, often detectable as smooth variations in the stellar density of old disk stars. A *flocculent* spiral is made of small pieces of spiral structure that appear sheared by differential rotation. The appearance of these features can be strongly affected by dust, such that at longer wavelengths a flocculent spiral may appear more grand design. Pseudorings can be thought of as variable pitch angle spirals which close on themselves, as opposed to continuously opening, constant pitch angle, logarithmic spirals.

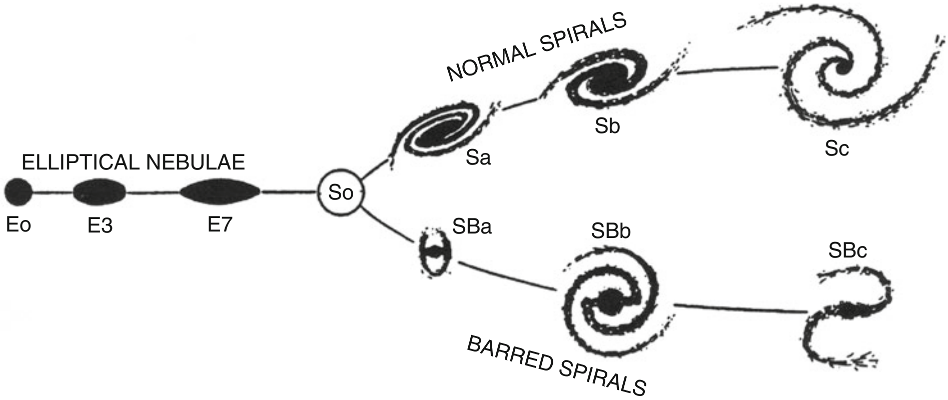
There are also numerous structures outside the scope of traditional galaxy classification, often connected with strong interactions between galaxies. Plus, the above described features are not necessarily applicable or relevant to what we see in very distant galaxies. Accounting for all of the observed features of nearby galaxies, and attempting to connect what we see in nearby to what is seen at high redshift, is a major goal of morphological studies.

3 Galaxy Classification

As noted by Sandage (1975), the first step in studying any class of objects is a classification of those objects. Classification built around small numbers of shared characteristics can be used for sorting galaxies into fundamental categories, which can then be the basis for further research. From such research, physical relationships between identified classes may emerge, and these relationships may foster a theoretical interpretation that places the whole class of objects into a broader context.

The basic idea of galaxy classification is to take the complex combinations of structures described in the previous section and summarize them with a few type symbols. Sandage (1975) describes the earlier classification systems of Wolf, Reynolds, Lundmark, and Shapley that fell into disuse more than 50 years ago. The Morgan (1958) spectral type/concentration classification system, which was based on a connection between morphology (specifically central concentration) and the stellar content of the central regions, was used recently by Bershady et al. (2000) in a mostly quantitative manner (see also Abraham et al. 2003). Thus, Morgan's system has in a way survived into the modern era but not in the purely visual form that he proposed. Only one Morgan galaxy type, the supergiant cD type, is still used extensively (➋ Sect. 10.6). Van den Bergh's luminosity/arm morphology classification system is described by van den Bergh (1998; see➋ Sect. 6.5).

The big survivor of the early visual classification systems was that of Hubble (1926, 1936), as later revised and expanded upon by Sandage (1961) and de Vaucouleurs (1959). Sandage (1975) has argued that one reason Hubble's view prevailed is that he did not try and account for every superficial detail but kept his classes broad enough that the vast majority of galaxies could be sorted into one of his proposed bins. These bins were schematically illustrated in Hubble's famous "tuning fork"[1] (Hubble 1936; reproduced in ➋ *Fig. 1-2*), recognizing a sequence of progressive flattenings from ellipticals to spirals. Ellipticals had only two classification details: the smoothly declining brightness distribution with no inflections and no evidence for a disk and the ellipticity of the isophotes, indicated by a number after the "E" symbol. (For example, E3 means the ellipticity is 0.3.) Spirals were systems more flattened than an E7 galaxy that could be subdivided according to the degree of central concentration, the degree of openness of the arms, the degree of resolution of the arms into complexes of star formation (all three criteria determine position along the fork), and the presence or absence of a bar (determining the appropriate prong of the fork).



◘ Fig. 1-2
Hubble's (1936) "tuning fork" of galaxy morphologies is the basis for modern galaxy classification

[1] As recently noted by D. L. Block (Block et al. 2004a), this diagram may have been inspired by a similar schematic by Jeans (1929).

The S0 class at the juncture of the prongs of the fork was still hypothetical in 1936. As "armless disk galaxies," S0s were mysterious because all examples known in 1936 were barred. These were classified as SBa, but this was a troubling inconsistency because nonbarred Sa galaxies had full spiral patterns. Hubble predicted the existence of nonbarred S0s to fill the gap between type E7[2] and Sa and cure what he felt was a "cataclysmic" transition.

It was not long before Hubble himself realized that the tuning fork could not adequately represent the full diversity of galaxy morphologies, and after 1936, he worked on a revision that included real examples of the sought group: nonbarred S0 galaxies. Based on fragmentary notes he left behind, Sandage (1961) prepared the *Hubble Atlas of Galaxies* to illustrate Hubble's revision and also added a third dimension: the presence or absence of a ring. This was the first major galaxy atlas illustrating a classification system in a detailed, sophisticated way with beautifully produced photographs. Hubble's revision, with van den Bergh luminosity classes (Sandage and Tammann 1981), was updated and extended to types later than Sc by Sandage and Bedke (1994).

Because Sandage (1961) and Sandage and Bedke (1994) describe the Hubble–Sandage revision so thoroughly, the details will not be repeated here. Instead, the focus of the next section will be on the de Vaucouleurs revision as outlined in the dVA. The reasons for this are (1) the de Vaucouleurs classification provides the most familiar galaxy types to extragalactic researchers, mostly because of extensive continuing use of the Third Reference Catalogue of Bright Galaxies (RC3, de Vaucouleurs et al. 1991) and (2) the de Vaucouleurs classification is still evolving to cover more details of galaxy morphology considered significant at this time. It should be noted that both the de Vaucouleurs and Hubble–Sandage revisions are strictly applicable only to $z \approx 0$ galaxies and that it is often difficult to fit objects having $z > 0.5$ neatly into the categories defining these classification systems. High redshift galaxy morphology is described in ➲ Sect. 13.

4 A Continuum of Galactic Forms

The Hubble tuning fork is useful because it provides a visual representation of information Hubble (1926) had only stated in words. The fork contains an implication of continuity. For example, it does not rule out that there might be galaxies intermediate in characteristics between an "Sa" or "Sb" spiral or between a normal "S" spiral and a barred "SB" spiral. Continuity along the elliptical galaxy sequence was always implied as a smooth variation from round ellipticals (E0) to the most flattened ellipticals (E7). Sandage (1961) describes the modification that made the Hubble system more three-dimensional: the introduction of the (r) (inner ring) and (s) (pure spiral) subtypes. Continuity between these characteristics was possible using the combined subtype (rs). Thus, already by 1961, the Hubble classification system had become much more complicated than it was in 1926 or 1936. The addition of the S0 class was one reason for this, but the (r) and (s) subtypes were another.

In the Hubble–Sandage classification, it became common to denote galaxies on the left part of the Hubble sequence as "early-type" galaxies and those on the right part as "late-type" galaxies. By the same token, Sa and SBa spirals became "early-type spirals" while Sc and SBc spirals became "late-type spirals." Sb and SBb types became known as "intermediate-type spirals." The reason for these terminologies was convenience and borrows terminology often used for stars.

[2] Van den Bergh (2009a) shows that E0–E4 galaxies are more luminous on average than are E5–E7 galaxies, suggesting that all E7 galaxies (and not many have been recognized) are actually S0 galaxies. Van den Bergh argues that genuine E galaxies may be no more flattened than E6.

Young, massive stars of spectral classes O and B were known as "early-type stars" while older stars of cooler spectral types were known as "late-type stars." Hubble stated that his use of these temporal descriptions for galaxies had no evolutionary implications. An irony in this is that it eventually became clear that early-type galaxies are dominated by late-type stars, while late-type galaxies often have significant numbers of early-type stars.

De Vaucouleurs (1959) took the idea of continuity of galaxy morphology a step further by developing what he referred to as the *classification volume* (➋ *Fig. 1-3*). In this revision of the Hubble–Sandage (1961) classification, galaxy morphology represents a continuous sequence of forms in a three-dimensional volume with a long axis and circular cross sections of varying size. The long axis of the volume is the *stage*, or type, and it represents the long axis of the original Hubble tuning fork. The short axes are the family and the variety, which refer to apparent bar strength and the presence or absence of an inner ring, respectively. In addition to Hubble's original stages E, Sa, Sb, and Sc, the classification volume includes new stages: late ellipticals, E^+; "very late" spirals, Sd; "Magellanic spirals," Sm; and "Magellanic irregulars," Im. The S0 class is included in the same position along the sequence, between E's and spirals, but is subdivided into three stages. The characteristics defining individual stages are described further in ➋ Sect. 5.2.

The stage is considered the most fundamental dimension of the classification volume because measured physical parameters, such as integrated color indices, mean surface brightnesses, and neutral hydrogen content, correlate well with position along the sequence (e.g., Buta et al. 1994). Early-type galaxies tend to have redder colors, higher average surface brightnesses, and lower neutral hydrogen content than late-type galaxies. The family and variety axes

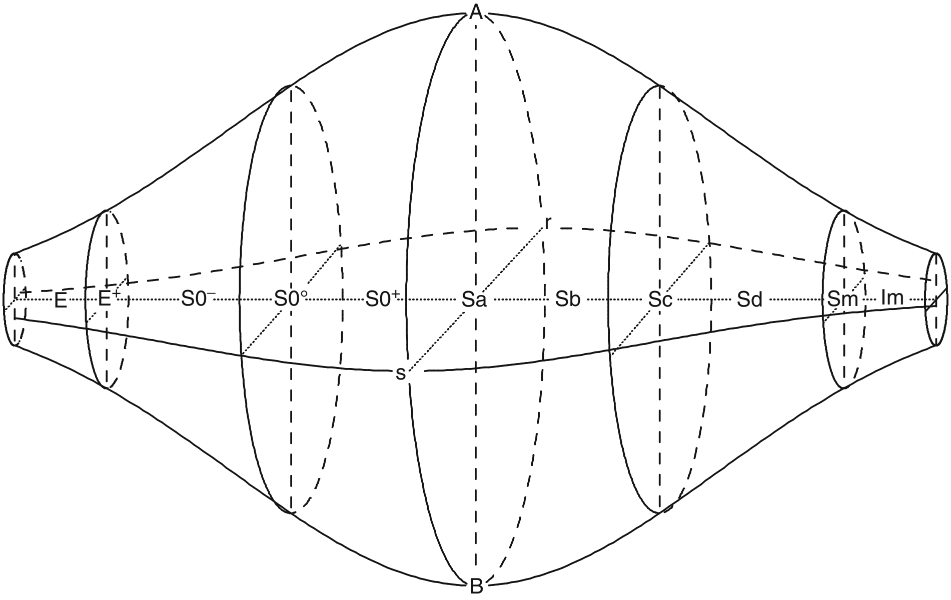


Fig. 1-3

de Vaucouleurs's (1959) classification volume, a revision and extension of the Hubble tuning fork. The three dimensions are the stage (Hubble type), the family (apparent bar strength), and the variety (presence or absence of an inner ring)

of the classification volume indicate the considerable variations in morphology at a given stage. A famous sketch of families and varieties near stage Sb, drawn by de Vaucouleurs himself during a cloudy night at McDonald Observatory circa 1962, is shown in Sandage (1975) and in Fig. 1.13 of the dVA. The classification volume is broader in the middle compared to the ends because the diversity of galaxy morphology is largest at stages like S0/a and Sa. Bars and rings are often most distinct and most recognizable at these stages. Such features are not characteristic of E galaxies, so the volume must be narrow at that end. Along the S0 sequence, bars and rings are barely developed among the earliest S0s ($S0^-$) and well developed among the late S0s ($S0^+$); thus, the volume begins to broaden. Among very late-type galaxies, Sd, Sdm, Sm, and Im, bars are actually very frequent, but closed rings (r) are not. Thus, the volume narrows at that end as well.

For the purposes of illustrating morphology, blue-light digital images converted to units of magnitudes per square arcsecond are used when available. This approach is described in the dVA, and requires calibration of the images, usually based on published photoelectric multiaperture photometry. In addition, some of the illustrations used (especially in ➋ Sect. 15) are from the Sloan Digital Sky Survey or from other sources. These are not in the same units but still provide excellent illustrations of morphology.

Unlike the dVA, the scope of this chapter extends beyond the traditional $UBVRI$ wavebands. It is only during the past 20 years that significant morphological information has been obtained for galaxies outside these bands, mostly at mid- and far-ultraviolet and mid-IR wavelengths from space-based observatories capable of imaging in these wavelengths to unprecedented depths, providing a new view of galaxy morphology that is only beginning to be explored. A useful review of many issues in morphology is provided by van den Bergh (1998).

5 Galaxy Types: Stage, Family, and Variety

5.1 Elliptical and Spheroidal Galaxies

Elliptical galaxies are smooth, amorphous systems with a continuously declining brightness distribution and no breaks, inflections, zones, or structures, as well as no sign of a disk. ➋ *Figure 1-4* shows some good examples. Because ellipticals are dominated by old stars and are relatively dust-free, they look much the same at different wavelengths. Hubble's subclassification of ellipticals according to apparent ellipticity (En, where $n = 10(1-b/a)$, b/a being the apparent flattening) was useful, but virtually no physical characteristics of ellipticals correlate with this parameter (Kormendy and Djorgovski 1989). The n value in the En classification is simply the projected ellipticity and not easily interpreted in terms of a true flattening without direct knowledge of the orientation of the symmetry planes. Luminous ellipticals are thought to be triaxial in structure with an anisotropic velocity dispersion tensor, while lower luminosity ellipticals are more isotropic oblate rotators (Davies et al. 1983). Studies of rotational to random kinetic energy (V/σ) versus apparent flattening (ϵ) show that massive ellipticals are slow anisotropic rotators. Ellipticals follow a fundamental plane relationship between the effective radius r_e of the light distribution, the central velocity dispersion σ_0, and mean effective surface brightness $< I_e >$ (see review by Kormendy and Djorgovski 1989). Dwarf elliptical galaxies may not follow the same relation as massive ellipticals; this is discussed by Ferguson and Binggeli (1994).

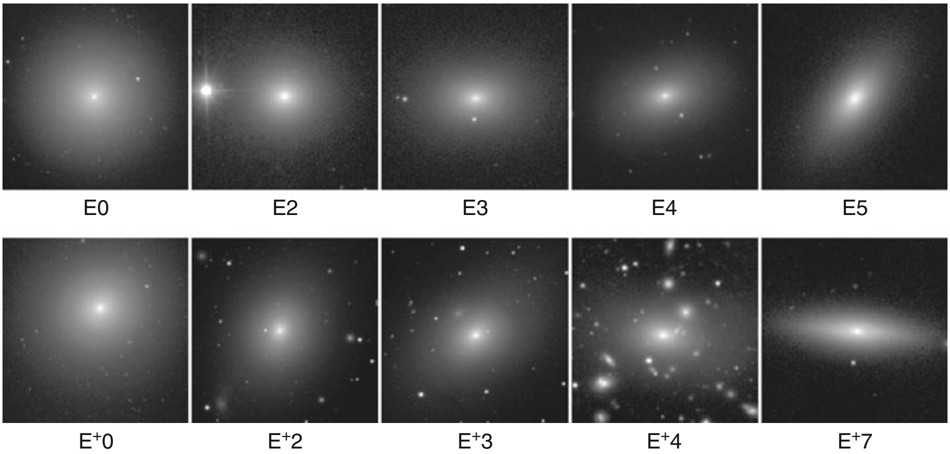


Fig. 1-4

Examples of elliptical galaxies of different projected shapes. Type E galaxies are normal ellipticals with no structural details. From *left* to *right* the galaxies shown are NGC 1379, 3193, 5322, 1426, and 720. Type E$^+$ galaxies are "late" ellipticals, which may include faint extended envelopes typical of large cluster ellipticals or simple transition types to S0$^-$. The examples shown are (*left* to *right*): NGC 1374, 4472, 4406, 4889, and 4623. All of these images are from the dVA (filters *B*, *V*, and *g*)

The lack of fundamental significance of Hubble's En classification led some authors to seek an alternative, more physically useful approach. Kormendy and Bender (1996; see also the review by Schweizer 1998) proposed a revision to Hubble's tuning fork handle that orders ellipticals according to their velocity anisotropy, since this is a significant determinant of E-galaxy intrinsic shapes. Velocity anisotropy correlates with the *deviations* of E galaxy isophotes from pure elliptical shapes, measured by the parameter a_4/a, the relative amplitude of the $\cos 4\theta$ Fourier term of these deviations. If this relative amplitude is positive, then the isophotes are pointy, disky ovals, while if negative, the isophotes are boxy ovals. A boxy elliptical is classified as E(b), while a disky elliptical is classified as E(d). The correlation with anisotropy is such that E(b) galaxies have less rotation on average and more velocity anisotropy than E(d) galaxies.

The classification proposed by Kormendy and Bender is shown in *Fig. 1-5*, together with two exceptional examples that show the characteristic isophote shapes. The leftmost of these two, NGC 7029, is an unusually obvious boxy elliptical at larger radii. This boxiness is the basis for the classification E(b)5. However, NGC 7029 is not boxy throughout: it shows evidence of a small inner disk and hence is disky at small radii. This is not necessarily taken into account in the classification. The other image shown in *Fig. 1-5* is of NGC 4697, a galaxy whose isophotes are visually disky. The idea with the Kormendy and Bender classification is that it is the disky ellipticals which connect to S0 galaxies and not the boxy ellipticals. However, NGC 7029 demonstrates that diskiness and boxiness can be a function of radius, thus perhaps a smooth connection between E(b) and E(d) types (i.e., type E(b,d)) is possible and the order shown in the Kormendy and Bender revision to the Hubble tuning fork, with boxy Es blending into the disky Es, could be reasonable.

Figure 1-5 shows two other galaxies that are clearly related to the boxy-disky ellipticals, only the inner disk is much more prominent, to the point that both galaxies are generally classified as S0s. The leftmost of these two, NGC 4638, has a bright edge-on disk which falls short of

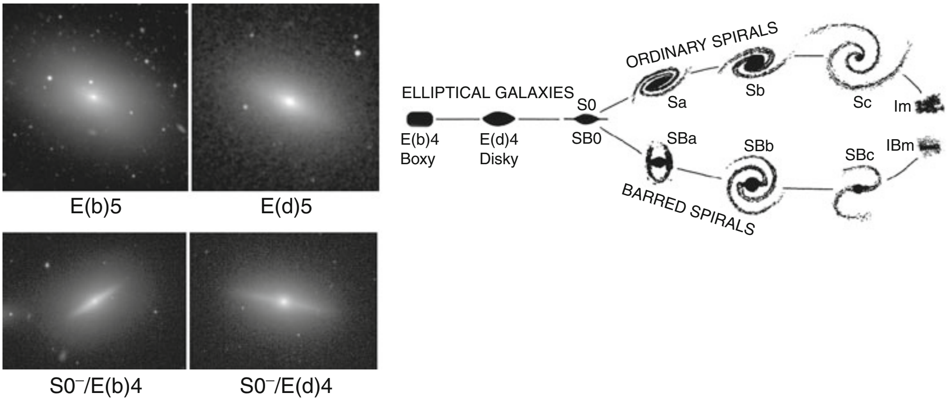


Fig. 1-5

Revised classification of elliptical galaxies from Kormendy and Bender (1996), as schematically incorporated into Hubble's (1936) "tuning fork." At *upper left* are two examples of boxy and disky ellipticals: NGC 7029 (*left*, *B*-band) and NGC 4697 (*JHK*$_s$ composite, 2MASS image from NED). At *lower left* are two related examples where the inner disk is much more prominent: NGC 4638 (*left*) and NGC 4474 (*right*). The images of these are SDSS *g*-band

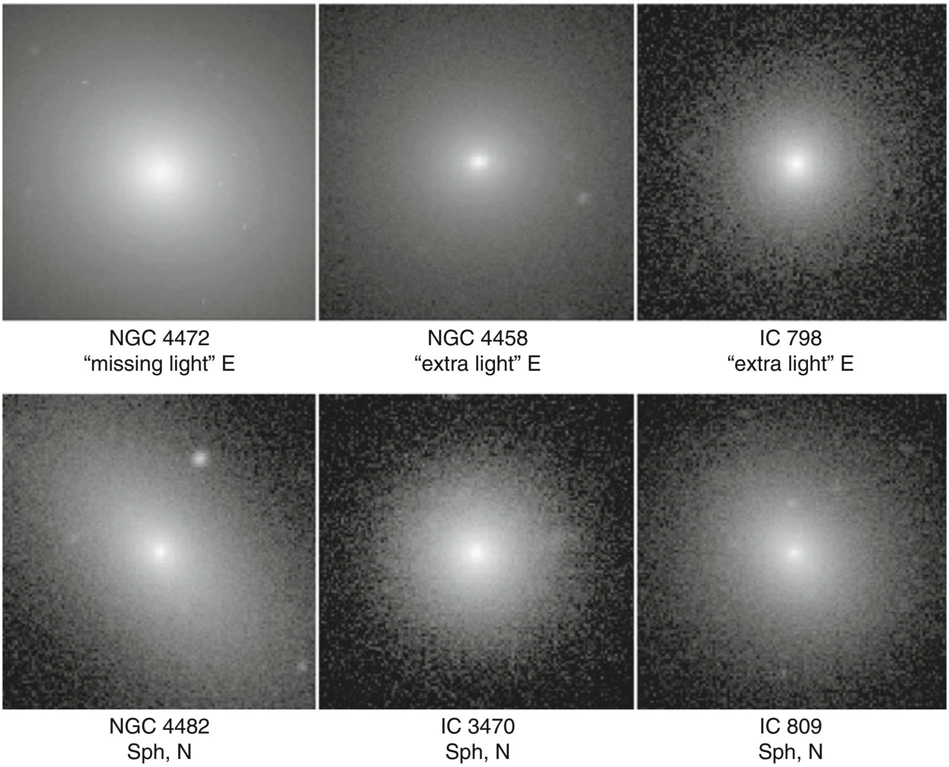
boxy outer isophotes, and is similar to NGC 7029. The rightmost, NGC 4474, is very much like NGC 4697 but again has a more prominent disk. Michard and Marchal (1993) describe cases like NGC 4638 and 4474 as showing "a disk fully embedded in a quasi-spherical envelope," while other S0s "remain pure disk, except for their central bulge." The classification S0$^-$/E(b) and S0$^-$/E(d) is proposed to recognize cases like NGC 4638 and 4474, respectively, using de Vaucouleurs's early S0 notation (Sect. 5.2).[3]

Note that while the classical En classifications of ellipticals were designed by Hubble to be estimated by eye, this is not easily done for the E(b) and E(d) classifications, which are most favored to be distinguished only when the disk is nearly edge-on. Face-on Es with imbedded disks will not show disky isophotes. For example, NGC 7029 and 4697 are extreme cases where the isophotal deviations are obvious by eye. But for most E galaxies, the E(b)n and E(d)n classifications can only be judged reliably with measurements of the a_4/a parameter.

The de Vaucouleurs classification of ellipticals includes a slightly more advanced type called E$^+$, or "late" ellipticals. It was originally intended to describe "the first stage of the transition to the S0 class" (de Vaucouleurs 1959). Five examples of E$^+$ galaxies are shown in the second row of *Fig. 1-4*. Galaxies classified as E$^+$ can be the most subtle S0s, but many of the E$^+$ cases listed in RC3 are the brightest members of clusters that have shallow enough brightness profiles to appear to have an extended envelope (see Sect. 10.6). Of the five E$^+$ galaxies shown, NGC 4623 (rare type E$^+$7 in the dVA) is of a much lower luminosity than the other four cases shown. While S0$^-$ is the type most often confused with ellipticals in visual classification, the bin has a wide spread from the most obvious to the least obvious cases. Thus, a type like E$^+$ is still useful for distinguishing transitions from E to S0 galaxies.

[3] After this chapter was completed, Kormendy and Bender (2012) suggested that NGC 4638 represents a harassed disk galaxy in a dense part of the Virgo Cluster and is not an E galaxy with an acquired or imbedded disk.

The photometric properties of ellipticals depend on luminosity. In terms of Sersic $r^{\frac{1}{n}}$ profile fits, large, luminous ellipticals tend to have profiles described better by $n \gtrsim 4$, while smaller, lower luminosity ellipticals tend to have $n < 4$, with values as low as 1 (e.g., Caon et al. 1993). Graham and Guzmán (2003) discuss the implications of this correlation on proposed dichotomies of elliptical galaxies (e.g., Kormendy 1985; Ferrarese et al. 2006). These studies received a major impetus from the massive photometric analysis of Virgo Cluster elliptical galaxies by Kormendy et al. (2009). Two issues were considered by these authors (*Fig. 1-6*). The first was whether galaxies classified as dwarf ellipticals ("dE"; see Sect. 15.2) in the Virgo Cluster really are the low-luminosity extension of more massive, conventional ellipticals or something different altogether. Based on parameter correlations, such as $r_{10\%}$, the major axis radius of the isophote containing 10% of the total visual luminosity, versus $\mu_{10\%}$, the surface brightness of this isophote, Kormendy et al. show that even the most elliptical-like and luminous dE galaxies lie on a distinct sequence from normal elliptical galaxies, which tend to lie on a


NGC 4472
"missing light" E

NGC 4458
"extra light" E

IC 798
"extra light" E

NGC 4482
Sph, N

IC 3470
Sph, N

IC 809
Sph, N

Fig. 1-6

Illustrations of six early-type galaxies in the Virgo Cluster with photometric classifications from Kormendy et al. (2009): NGC 4472 (core E, $M_V = -23.2$), NGC 4458 (coreless E, $M_V = -19.0$), IC 798 (VCC 1440; "coreless" E, $M_V = -16.9$), NGC 4482 (nucleated spheroidal, $M_V = -18.4$); IC 3470 (VCC 1431; nucleated spheroidal, $M_V = -17.4$), IC 809 (VCC 1910; nucleated spheroidal, $M_V = -17.4$). The images shown are all based on SDSS g-band single or mosaic images and are in units of mag arcsec^{-2}

higher density sequence. In a graph of B-band central surface brightness versus absolute B-band magnitude, the dE galaxies lie in a region occupied by Magellanic irregular galaxies, suggesting a link between the groups and consistent with the earlier conclusions of Kormendy (1985). Kormendy et al. (2009) suggest reclassifying Binggeli et al. (1985) "dE" galaxies and related objects (like dwarf S0 or dS0 galaxies) as "spheroidal" (Sph) galaxies, including the type "Sph,N," meaning "nucleated spheroidal" galaxy (➜ Sect. 15.2). ➜ *Figure 1-6* shows several examples of Sph,N galaxies as compared with several genuine elliptical galaxies. The morphological appearance alone does not necessarily distinguish the two classes. The classification is physical, being based mainly on parameter correlations. Kormendy et al. suggest that Sph galaxies are formed from late-type systems by environmental effects and supernovae. The "home" for Sph galaxies in the Hubble tuning fork classification is described further in ➜ Sect. 17.

The second issue considered was the physical distinction between "core" elliptical galaxies, those where the surface brightness profile approaches either a constant level or a slightly sloped level with radius approaching zero, and "coreless" ellipticals (also known as "power law" Es) where the inner profile steepens with decreasing radius (Kormendy 1999). Kormendy et al. (2009) illustrated both types relative to a Sersic $r^{\frac{1}{n}}$ fit to the outer regions of the luminosity profiles. In this representation, core Es are "missing light" relative to the fit while coreless Es have "extra light." The top row of ➜ *Fig. 1-6* shows one core E (NGC 4472) and two coreless Es (NGC 4458 and IC 798 (VCC 1440), the latter a low-luminosity dwarf). The subtle distinctions are evident in these images, with NGC 4472 showing a soft center and NGC 4458 showing a strong center. The terminology for both types is mostly historical (Kormendy 1999) and somewhat counter to the visual impression (i.e., NGC 4472 lacks a bright central peak while NGC 4458 has one, yet the latter is technically coreless). Kormendy et al. show that core and coreless E galaxies have different Sersic indices, velocity dispersion anisotropy, isophote shapes, and rotational character, with the core Es being of the boxy type and the coreless Es of the disky type in ➜ *Fig. 1-5*. The distinction may be tied to the number of mergers that formed the system.

5.2 S0 and Spiral Galaxies

The full classification of spiral and S0 galaxies involves the recognition of the stage, family, and variety. In de Vaucouleurs's classification approach, the implication for bars, inner rings, and stages is a continuum of forms (de Vaucouleurs 1959), so that there are no sharp edges to any category or "cell" apart from the obvious ones (e.g., there are no galaxies less "barred" than a nonbarred galaxy, nor are there galaxies more ringed than those with a perfectly closed ring).

The classification of S0 galaxies depends on recognizing the presence of a disk and a bulge at minimum, and usually a lens as well, and no spiral arms. Examples are shown in ➜ *Fig. 1-7*. The display of galaxy images in units of mag arcsec^{-2} makes lenses especially easy to detect, as noted in the dVA. Even if a lens is not obvious, a galaxy could still be an S0 if it shows evidence of an exponential disk. (Lenses are also not exclusive to S0s.) The "no spiral arms" characteristic is much stricter in the Hubble–Sandage classification than in the de Vaucouleurs interpretation because varieties (r, rs, and s) are carried into the de Vaucouleurs classifications of S0s. This allows the possibility of a classification like SA(s)0^{-}, which would be very difficult to recognize. Bars enter in the classification of S0s in a similar manner as that of spirals. ➜ *Figure 1-7* shows mainly stage differences among nonbarred and barred S0s. The stage for S0s ranges from early (S0^{-}) to intermediate (S0°), and finally to late (S0^{+}), in a succession of increasing detail. The earliest nonbarred S0s may be mistaken for elliptical galaxies on photographic images, and indeed

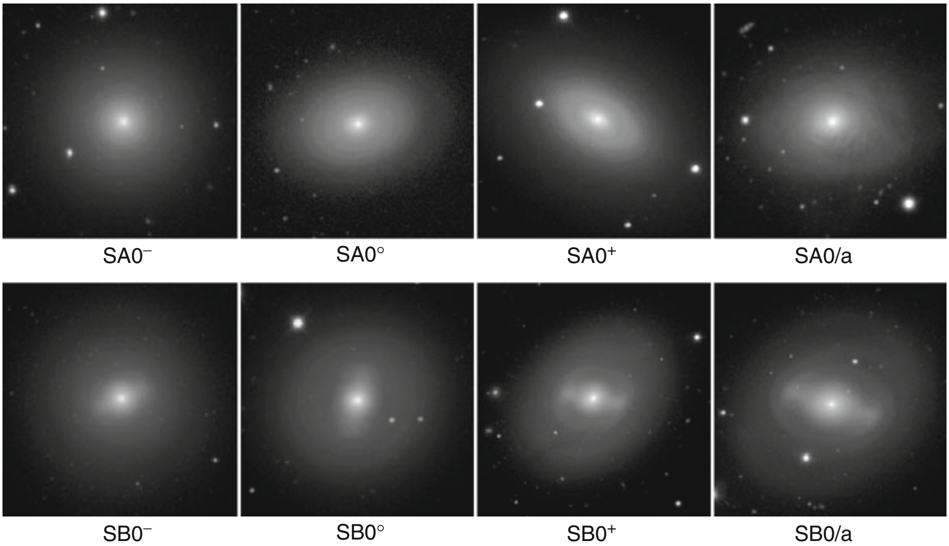


Fig. 1-7
Examples of barred and nonbarred S0 galaxies of different stages from "early" (S0$^-$) to "intermediate" (S0°) to "late" (S0$^+$), including the transition stage to spirals, S0/a. The galaxies shown are (*left* to *right*): *Row 1* – NGC 7192, 1411, 1553, and 7377; *Row 2* – NGC 1387, 1533, 936, and 4596. All images are from the dVA (filters *B* and *V*)

Sandage and Bedke (1994) note cases where they believe an S0 galaxy has been misclassified as an elliptical by de Vaucouleurs in his reference catalogues (see also *The Revised Shapley-Ames Catalogue*, RSA, Sandage and Tammann 1981). This kind of misinterpretation is less likely for types S0° and S0$^+$ because these will tend to show more obvious structure.

The morphological distinction between E and S0 galaxies has been considered from a quantitative kinematic point of view by Emsellem et al. (2007). These authors argue that the division of early-type galaxies into E and S0 types is "contrived" and that it is more meaningful to divide them according to a quantitative kinematic parameter called λ_R, the specific angular momentum of the stellar component, which is derived from a two-dimensional velocity field obtained with the SAURON integral field spectrograph (Bacon et al. 2001). Based on this parameter, early-type galaxies are divided into slow and fast rotators, that is, whether they are characterized by large-scale rotation or not. In a sample of 48 early-types, most were found to be fast rotators classified as a mix of E and S0 types, while the remainder were found to be slow rotators classified as Es. This kind of approach, which provides a more physical distinction among early-types, does not mean the classical E and S0 subdivisions no longer have value but highlights again the persistent difficulty of distinguishing the earliest S0s from Es by morphology alone.

The transition type S0/a shows the beginnings of spiral structure. Two examples are included in *Fig. 1-7*, one nonbarred and the other barred. Type S0/a is a well-defined stage characterized in the de Vaucouleurs 3D classification volume as having a high diversity in family and variety characteristics. The type received a negative characterization as the "garbage bin" of the Hubble sequence at one time because troublesome dusty irregulars, those originally classified as "Irr II" by Holmberg (1950) and later as "I0" by de Vaucouleurs, seemed to fit better in that part

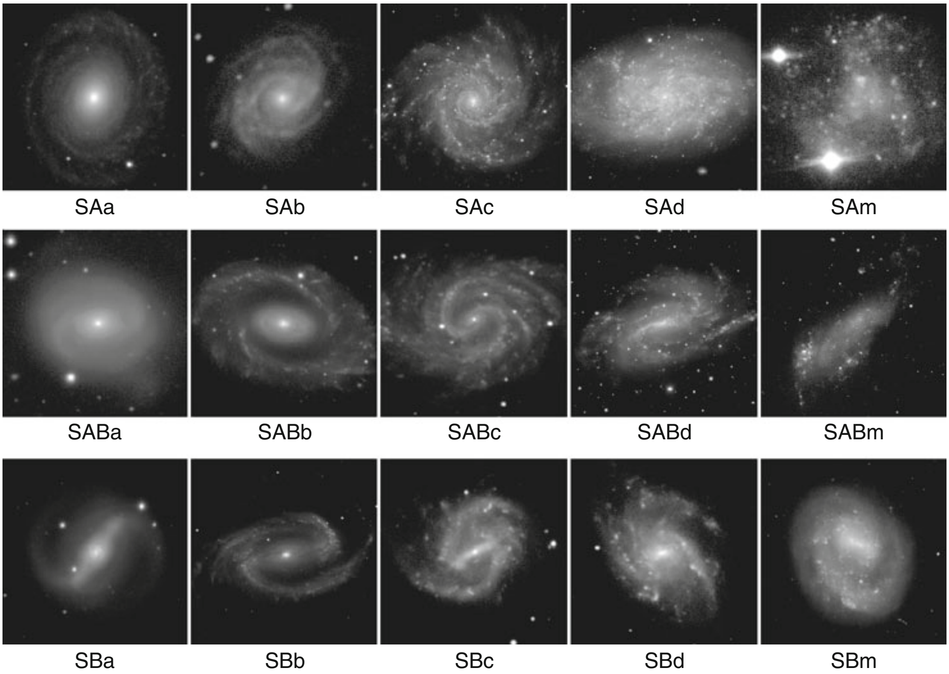


Fig. 1-8

Stage classifications for spirals divided according to bar classifications into parallel sequences. The galaxies illustrated are (*left* to *right*): *Row 1* – NGC 4378, 7042, 628, 7793, and IC 4182; *Row 2* – NGC 7743, 210, 4535, 925, and IC 2574; *Row 3* – NGC 4314, 1300, 3513, 4519, and 4618. All images are *B*-band from the dVA

of the sequence. (In fact, de Vaucouleurs et al. (1976) assigned the numerical stage index $T = 0$ to both S0/a and I0 galaxies.) However, this problem is only a problem at optical wavelengths. At longer wavelengths (e.g., 3.6 μm), types such as Irr II or I0 are less needed because they can be defined mainly by dust (Buta et al. 2010a).

In general, the stage for spirals is based on the appearance of the spiral arms (degree of openness and resolution) and also on the relative prominence of the bulge or central concentration. These are the usual criteria originally applied by Hubble (1926, 1936). *Figure 1-8* shows the stage sequence for spirals divided according to bar classification (SA, SAB, SB) and as modified and extended by de Vaucouleurs (1959) to include Sd and Sm types. Intermediate stages, such as Sab, Sbc, Scd, and Sdm, are shown in *Fig. 1-9*. As noted by de Vaucouleurs (1963), these latter stages are almost as common as the basic ones.

The three Hubble criteria are basically seen in the illustrated galaxies. Sa galaxies tend to have significant bulges and tightly wrapped and relatively smooth spiral arms. Sab galaxies are similar to Sa galaxies but show more obvious resolution of the arms. Sb galaxies have more resolution and more open arms, and generally smaller bulges than Sab galaxies. Sbc galaxies have considerable resolution and openness of the arms, and also usually significant bulges. In Sc galaxies, the bulge tends to be very small and the arms patchy and open. Scd galaxies tend to be relatively bulgeless, patchy armed Sc galaxies. Stage Sd is distinctive mainly as almost completely bulgeless late-type spirals with often ill-defined spiral structure.

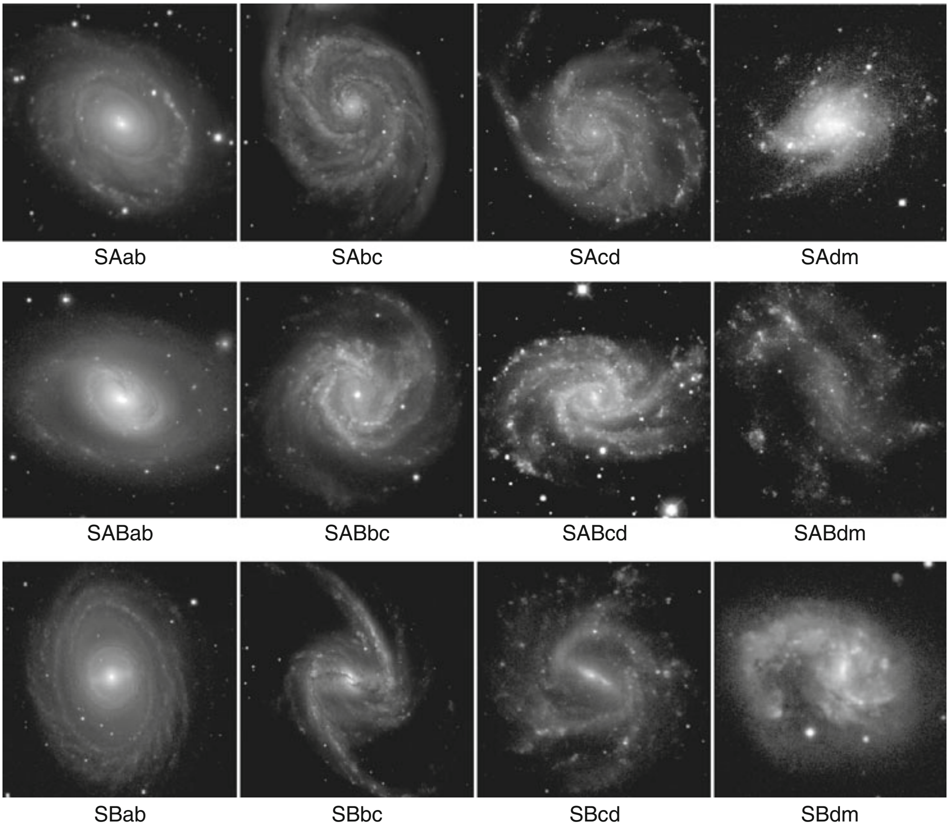


Fig. 1-9

Sequences of stages intermediate between the main stages illustrated in *Fig. 1-8*. The galaxies are (*left* to *right*): *Row 1* – NGC 2196, 5194, 5457, and 4534; *Row 2* – NGC 3368, 4303, 2835, and 4395; *Row 3* – NGC 1398, 1365, 1073, and 4027. All images are *B*-band from the dVA, except for NGC 4534, which is SDSS *g*-band

Stages Sdm and Sm are the most characteristically asymmetric stages, the latest spiral types along the de Vaucouleurs revised Hubble sequence. They are described in detail by de Vaucouleurs and Freeman (1972) and by Odewahn (1991). Sm is generally characterized by virtually no bulge and a single principle spiral arm. If a bar is present, it is usually not at the center of the disk isophotes, unlike what is normally seen in earlier type barred spirals. This leads to the concept of an *offset barred galaxy*. The single spiral arm emanates from one end of the bar. As noted by Freeman (1975), this is a basic and characteristic asymmetry of the mass distribution of Magellanic barred spirals. Sdm galaxies are similar but may show a weaker or shorter second arm. In *Fig. 1-8*, NGC 4618 is an especially good example of an SBm type (Odewahn 1991), while in *Fig. 1-9*, NGC 4027 is illustrated as type SBdm.

An important issue regarding these galaxies is whether the optically offset bar is also offset from the dynamical rotation center of the disk. In a detailed HI study of the interacting galaxy pair NGC 4618 and 4625, Bush and Wilcots (2004) found very regular velocity fields and extended HI disks but no strong offset of the rotation center from the center of the bar.

This is similar to what Pence et al. (1988) found for the offset barred galaxy NGC 4027 based on optical Fabry–Perot interferometry. In contrast, both Magellanic Clouds, which are also offset barred galaxies, were found to have HI rotation centers significantly offset from the center of the bar (Kerr and de Vaucouleurs 1955).

In general, the application of Hubble's three spiral criteria allows consistent classification of spiral types. Nevertheless, sometimes the criteria are inconsistent. For example, small-bulge Sa galaxies are described by Sandage (1961) and Sandage and Bedke (1994). Barred galaxies with nuclear rings can have spiral arms like those of an earlier Hubble type and very small bulges. In such conflicting cases, the emphasis is usually placed on the appearance of the arms. Also, while late-type Sdm and Sm galaxies are characteristically asymmetric, other types may be asymmetric as well. On average, the bulge-to-total luminosity ratio is related to Hubble type, but the result is sensitive to how galaxies are decomposed (e.g., Laurikainen et al. 2005). Asymmetry has been quantified by Conselice (1997).

The family classifications SA, SAB, and SB are purely visual estimates of bar strength for both spirals and S0s. They are highlighted already in *Figs. 1-7*–1-9, but the continuity of this characteristic is better illustrated in *Fig. 1-10*, where de Vaucouleurs (1963) underlined classifications (SA̲B and SAB̲) are also shown. An SA galaxy has no evident bar in general, although high inclination can cause a mistaken SA classification if a bar is highly foreshortened. Also, internal dust may obscure a bar (see, e.g., Eskridge et al. 2000). An SB galaxy should have a clear, well-defined bar. The intermediate bar classification SAB is one of the hallmarks of the de Vaucouleurs system and is used to recognize galaxies having characteristics intermediate between barred and nonbarred galaxies. It is used for well-defined ovals or simply weaker-looking normal bars. The weakest primary bars are denoted SA̲B while the classification SAB̲ is usually assigned to more classical bars that appear only somewhat weaker than conventional bars. Most of the time, galaxies which should be classified as SAB̲ are simply classified as SB.

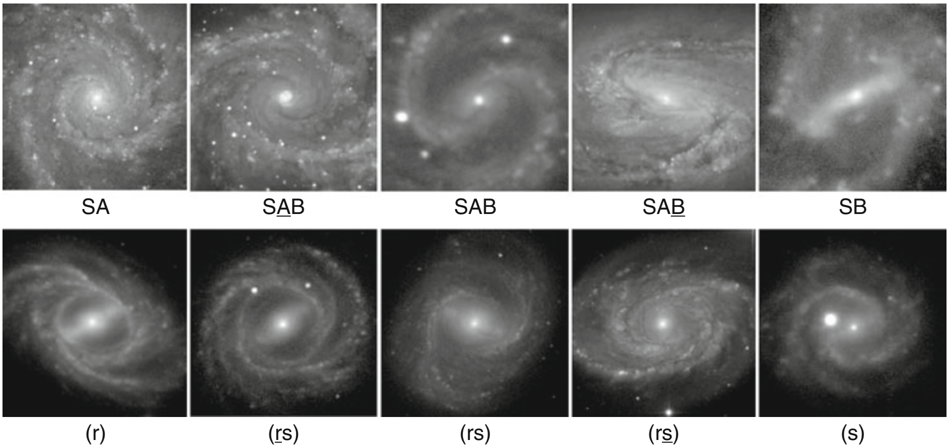


Fig. 1-10
The continuity of family and variety characteristics among spiral galaxies, including underlined classifications used by de Vaucouleurs (1963). The galaxies are (*left* to *right*): *Row 1* – NGC 628, 2997, 4535, 3627, and 3513; *Row 2* – NGC 2523, 3450, 4548, 5371, and 3507. All images are *B*-band from the dVA

Variety is also treated as a continuous classification parameter (*Fig. 1-10*, second row). A spiral galaxy having a completely closed or very nearly completely closed inner ring is denoted (r). The spiral arms usually break from the ring. If the spiral arms break directly from the central region or the ends of a bar, forming a continuously winding, open pattern, the variety is (s). The intermediate variety (rs) is also well defined. Inner rings which appear broken or partial are in this category. The "dash-dot-dash in brackets" morphology, (-o-), where a bar with a bulge is bracketed by spiral arcs overshooting the bar axis, is very typical of variety (rs). The example of this shown in *Fig. 1-10* is NGC 4548. We use the notation r̲s to denote an inner ring made up of tightly wrapped spiral arms that do not quite close, while the notation rs̲ is used for very open, barely recognizable, inner pseudorings. A good example of the former is NGC 3450, while an example of the latter is NGC 5371.

A *spindle* is a highly inclined disk galaxy. For blue-light images, usually an "sp" after the classification automatically implies considerable uncertainty in the interpretation because family and variety are not easily distinguished when the inclination is high. *Figure 1-11* shows, however, that stages can be judged reasonably reliably for edge-on galaxies. One important development in the classification of edge-on galaxies has been the ability to recognize edge-on bars through boxy/peanut and "X" shapes. Boxy/peanut bulges in edge-on galaxies were proven to be bars seen edge-on from kinematic considerations (e.g., Kuijken and Merrifield 1995). This shape is evident in NGC 4425 (row 1, column 4 of *Fig. 1-11*; see also Sect. 9).

For spiral and S0 galaxies that are not too highly inclined (i.e., not spindles), once the stage, family, and variety are determined, these are combined in the order family, variety, stage for a final full type. For example, NGC 1300 is of the family SB, variety (s), and stage b; thus, its full type is SB(s)b. The S0$^+$ galaxy NGC 4340 has both a bar and inner ring, and its full type is SB(r)0$^+$. The classification is flexible enough that if, for example, the family and variety of a galaxy cannot be reliably determined owing to high inclination, while the stage can still be assessed, then the symbols can be dropped and a type such as "Sb" or "S0" can still be noted.

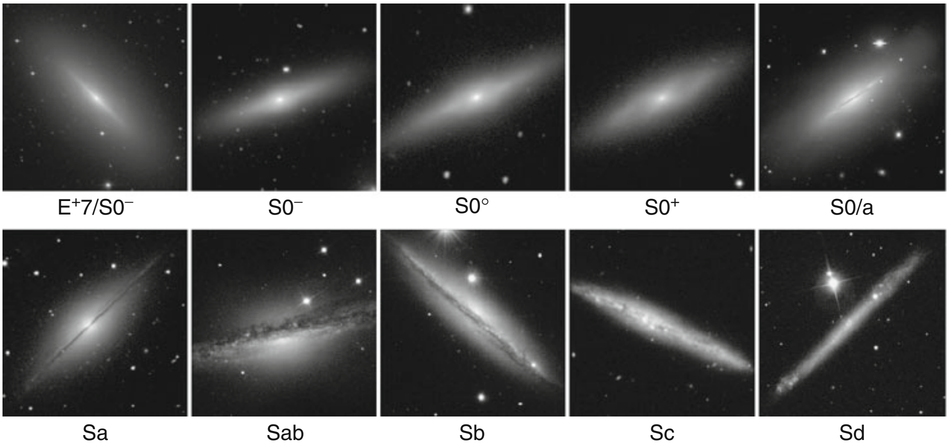


Fig. 1-11
Classification of edge-on galaxies by stage. The galaxies are (*left* to *right*): *Row 1* – NGC 3115, 1596, 7332, 4425, and 5866; *Row 2* – NGC 7814, 1055, 4217, 4010, and IC 2233. All images are from the dVA (*B* and *V* filters)

5.3 Irregular Galaxies

Magellanic irregular galaxies represent the last normal stage of the de Vaucouleurs revised Hubble sequence. Several examples are shown in *Fig. 1-12*. The objects illustrated in the top row are all examples of (s)-variety irregulars with bars or some trace of a bar. Nevertheless, not all Magellanic irregulars have bars. Irregulars of the lowest luminosities are usually classified simply as Im since the sophistication of structure needed to distinguish something like "family" may not exist for such galaxies.

Irregular galaxies are important for their star formation characteristics. As noted by Hunter (1997), irregulars are similar to spirals in having both old and young stars, as well as dust, atomic, molecular, and ionized gas, but lack the spiral structure that might trigger star formation. Thus, they are useful laboratories for examining how star formation occurs in the absence of spiral arms.

Although irregulars are largely defined by a lack of well-organized structure like spiral arms, the two lower right galaxies in *Fig. 1-12* are not so disorganized looking and seem different from the other cases shown. NGC 5253 looks almost like a tilted S0 galaxy, yet it has no bulge at its center, nor any obvious lenses. Instead, the central area is an irregular zone of active star formation. The central zone was interpreted by van den Bergh (1980a) as "fossil evidence" for a burst of star formation, possibly triggered by an interaction with neighboring M83. This is a case where the de Vaucouleurs classification of I0 seems reasonable: NGC 5253 is an early-type galaxy with a central starburst, probably the youngest and closest example known (Vanzi and Sauvage 2004). It is a Magellanic irregular galaxy imbedded in a smooth S0-like background known to have an early-type star spectrum. NGC 1705, also shown in *Fig. 1-12*, is similar but has a super star cluster near the center and obvious peculiar filaments. It is classified as a blue compact dwarf by Gil de Paz et al. (2003). Both galaxies are classified as amorphous by Sandage and Bedke (1994).

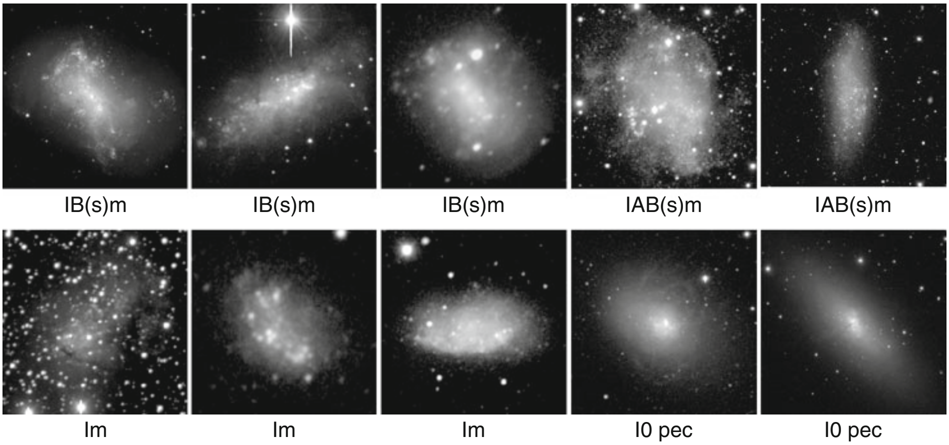


Fig. 1-12
Examples of irregular galaxies ranging in absolute blue magnitude from −14 to −18. The galaxies are (*left* to *right*): *Row 1* – NGC 4449, 1569, 1156, DDO 50, and A2359-15 (WLM galaxy); *Row 2* – IC 10, DDO 155, DDO 165, NGC 1705, NGC 5253. The "pec" stands for peculiar. All images are *B*-band from the dVA

6 Other Dimensions to Galaxy Morphology

The de Vaucouleurs classification volume recognizes three principal aspects of galaxy morphology, but there are many more dimensions than three. Stage, family, and variety are the dimensions most clearly highlighted in blue-light images and have a wide scope. Other dimensions may be considered, and for some, there is explicit notation in use.

6.1 Outer Rings and Pseudorings

Published de Vaucouleurs types include an extra dimension known as the outer ring/pseudoring classification. Several examples of outer rings and pseudorings are shown in *Fig. 1-13*. An outer ring is a large, often diffuse structure, typically seen in barred early-type galaxies (stages S0^{+} to Sa) at a radius approximately twice that of the bar. Closed outer rings are recognized with the type symbol (R) preceding the main part of the classification. For example, an SB(r)0^{+} galaxy having an outer ring has a full classification of (R)SB(r)0^{+}. Interestingly, rare cases of double outer ring galaxies, type (RR), are known, where two detached outer rings are seen; an example is NGC 2273 shown in the upper right frame of *Fig. 1-13*.

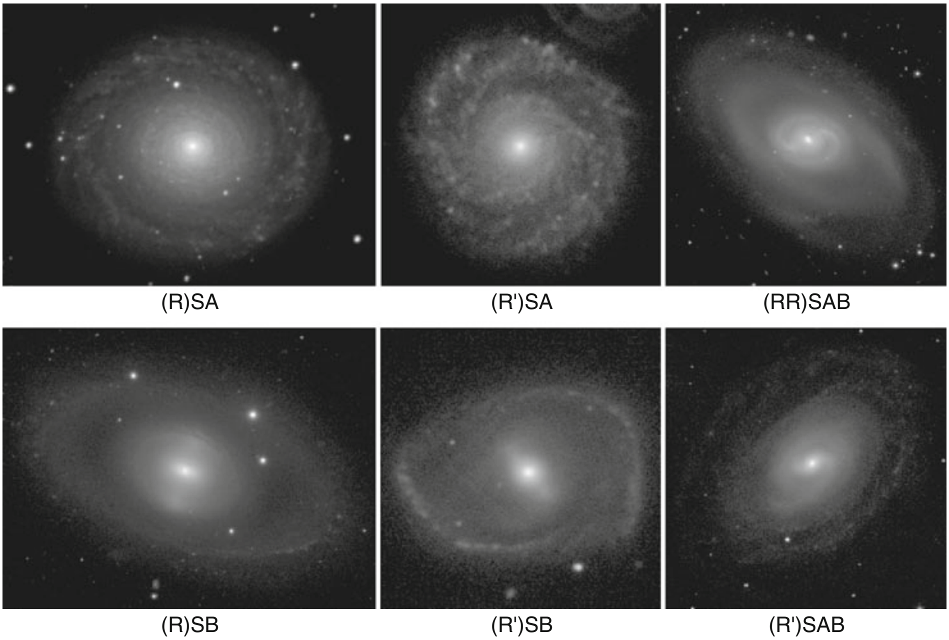


Fig. 1-13

Examples of outer rings (R) and outer pseudorings (R′) in barred and nonbarred galaxies. Also shown is a rare example with two largely detached outer rings (RR). The galaxies are (*left* to *right*): *Row 1* – NGC 7217, IC 1993, and NGC 2273; *Row 2* – NGC 3945, NGC 1358, and NGC 1371. All images are from the dVA and are *B*-band except for NGC 2273, which is *r*-band

In later-type galaxies, a large outer ringlike feature is often seen made of outer spiral arms whose variable pitch angle causes them to close together. These features are classified as outer pseudorings, symbolized by (R′) preceding the main type symbols (e.g., as in (R′)SB(r)ab). Outer pseudorings are mainly observed in Sa to Sbc galaxies and are only rarely seen in the very late stages Sc–Sm.

Among bright nearby galaxies, outer rings and pseudorings are found at about the 10% level (Buta and Combes 1996). Typically, outer rings are fainter than 24 mag arcsec^{-2} in blue light. With such low surface brightnesses, the rings can be easily lost to Galactic extinction. The division between outer rings and pseudorings is also not sharp. Some outer pseudorings are only barely distinguishable from outer rings. Continuity applies to these features as it does for inner rings although there is no symbol other than "S" for an outer spiral pattern which does not close into an outer pseudoring.

Although closed outer rings (R) are equally well recognized in the RSA and the Carnegie and Hubble Atlases, outer pseudorings are a unique feature of the de Vaucouleurs revision. The value of recognizing these features is that many show morphologies consistent with the theoretical expectations of the outer Lindblad resonance (OLR, Schwarz 1981), one of the major low-order resonances that can play a role in disk evolution. Resonance rings are discussed further in ➲ Sect. 10.1, but ➲ *Fig. 1-14* shows schematics of the morphologies generally linked to this resonance. The schematics are designed to highlight the subtle but well-defined aspects of these features, while ➲ *Fig. 1-15* shows images of several examples of each morphology, including the "models" used for the schematic. Outer rings of type R$_1$ are closed rings that are slightly dimpled toward the bar axis, a shape which connects directly to one of the main periodic orbit

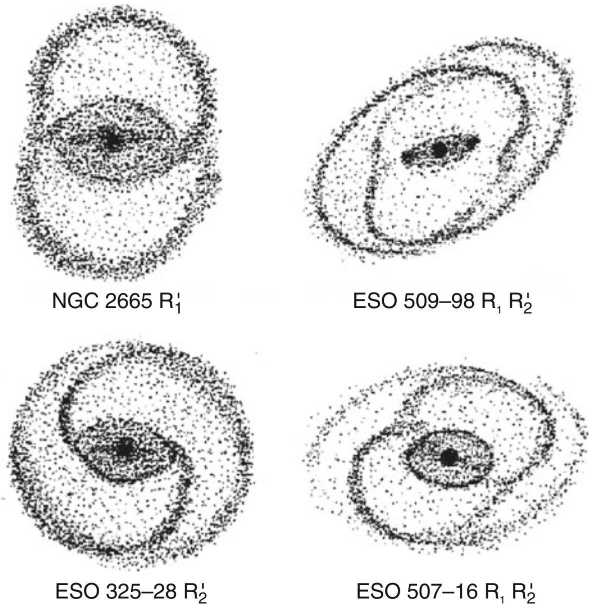


Fig. 1-14

Schematic representations of outer Lindblad resonance (OLR) morphologies (Buta and Combes 1996)

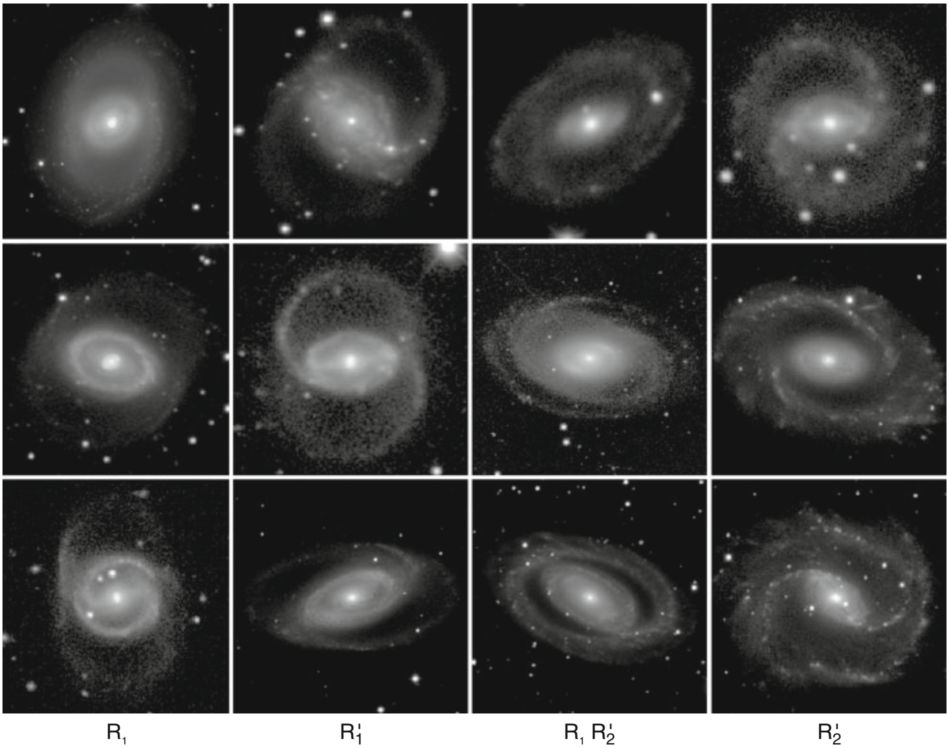


Fig. 1-15
Examples of OLR subclasses of outer rings and pseudorings. The galaxies are (*left* to *right*): *Row 1* – NGC 1326, NGC 2665, ESO 509–98, and ESO 325–28; *Row 2* – NGC 3081, UGC 12646, NGC 1079, and NGC 210; *Row 3* – NGC 5945, 1350, 7098, and 2935. All images are *B*-band from the dVA

families near the OLR as shown in Schwarz (1981) and in the dVA. Outer pseudorings of type R'_1 are similar to type R_1 but are made of two spiral arms that wind approximately 180° with respect to the ends of the bar. Even these will usually show a dimpled shape. Outer pseudorings of type R'_2 are different from this in that two spiral arms wind 270° with respect to the ends of the bar, such that the arms are doubled in the quadrants immediately trailing the bar.

The shapes R_1, R'_1, and R'_2 were predicted by Schwarz (1981) based on "sticky-particle" numerical simulations. Not predicted by those simulations (but later shown in extensions of those simulations by Byrd et al. 1994 and Rautiainen and Salo 2000) is an interesting combined ring morphology called $R_1R'_2$, which consists of a closed R_1 ring and an R'_2 pseudoring. This combination is especially important because it demonstrates not only a continuity of morphologies among outer rings and pseudorings different from the continuity between outer rings and pseudorings in general, but it is also a morphology that can be linked directly to the dynamics of barred galaxies.

Note that the classification shown in *Figs. 1-14* and *1-15* does not depend on whether the rings are in fact linked to the OLR. The illustrated morphologies are abundant and easily recognized regardless of how they are interpreted. (Sect. 10.1 discusses other interpretations that have been proposed.) Although the Schwarz models guided the search for these morphologies, Rautiainen et al. (2004) and Treuthardt et al. (2008) showed that some outer pseudorings classified as R_1' are more likely related to the outer 4:1 resonance and not the OLR. These cases are generally recognizable by the presence of secondary spiral arcs in a four-armed pattern in the area of the bar (NGC 1433 in ❯ *Fig. 1-1* and ESO 566–24 in ❯ *Fig. 1-19* are examples).

The OLR subclassifications are used in the same manner as the plain outer ring and pseudoring classifications. For example, NGC 3081 has the full type (R_1R_2')SAB(r)0/a.

6.2 Inner and Outer Lenses

The value of recognizing lenses as significant morphological components was first emphasized by Kormendy (1979), who suggested a dynamical link between inner lenses, which are often filled by a bar in one dimension, and dissolved or dissolving bars. Kormendy noted that lenses can be of the inner or outer type in a manner analogous to inner and outer rings. He suggested the notation (l) for inner lenses and (L) for outer lenses to be used in the same position of the classification as inner and outer rings. For example, the galaxy NGC 1543 is type (R)SB(l)0/a while galaxy NGC 2983 is type (L)SB(s)0^+. ❯ *Figure 1-16* demonstrates the continuity between rings and lenses, which is evident not only among barred galaxies but among nonbarred ones as well. This continuity is recognized by the type symbol (rl), also used by Kormendy (1979). This type refers to a low-contrast inner ring at the edge of a clear lens. Even underlined classifications ($\underline{\text{r}}$l) and (r$\underline{\text{l}}$) may be recognized. A rare classification, (r′l), refers to an inner pseudoring/lens, a type of feature that is seen in NGC 4314 and recognized as such in the dVA.

Similarly, ❯ *Fig. 1-17* shows a continuity between outer rings and outer lenses through the type classification (RL), referring to an outer lens with a weak ringlike enhancement. Underlined types $\underline{\text{R}}$L and R$\underline{\text{L}}$ may also be recognized. The origin of outer lenses could be in highly evolved outer rings.

6.3 Nuclear Rings and Bars

The central regions of barred galaxies often include distinct morphological features in the form of small rings and secondary bars. The rings, known as nuclear rings because of their proximity to the nucleus well inside the ends of the primary bar, are sites of some of the most spectacular starbursts known in normal galaxies. The rings are typically ≈1.5 kpc in linear diameter and intrinsically circular in shape. ❯ *Figure 1-18* (top row) shows three examples: NGC 1097, 3351, and 4314. These images highlight the small bulges that seem characteristic of nuclear-ringed barred galaxies. The three galaxies illustrated have types ranging from Sa to Sb, but based on the bulge size, the types would be considerably later. For example, NGC 3351 has the bulge of an Sd galaxy.

Comerón et al. (2010) carried out an extensive statistical study of nuclear ring radii and identified a subclass known as "ultracompact" nuclear rings (UCNRs). Such rings were recognized mainly in Hubble Space Telescope images and are defined to be less than 200 pc in diameter. (See ❯ *Fig. 1-26* for an example in NGC 3177.) Comerón et al. showed that UCNRs are the low size tail of the global nuclear ring population. This study also showed that bar strength impacts the sizes of nuclear rings, with stronger bars generally hosting smaller nuclear rings than weaker bars.

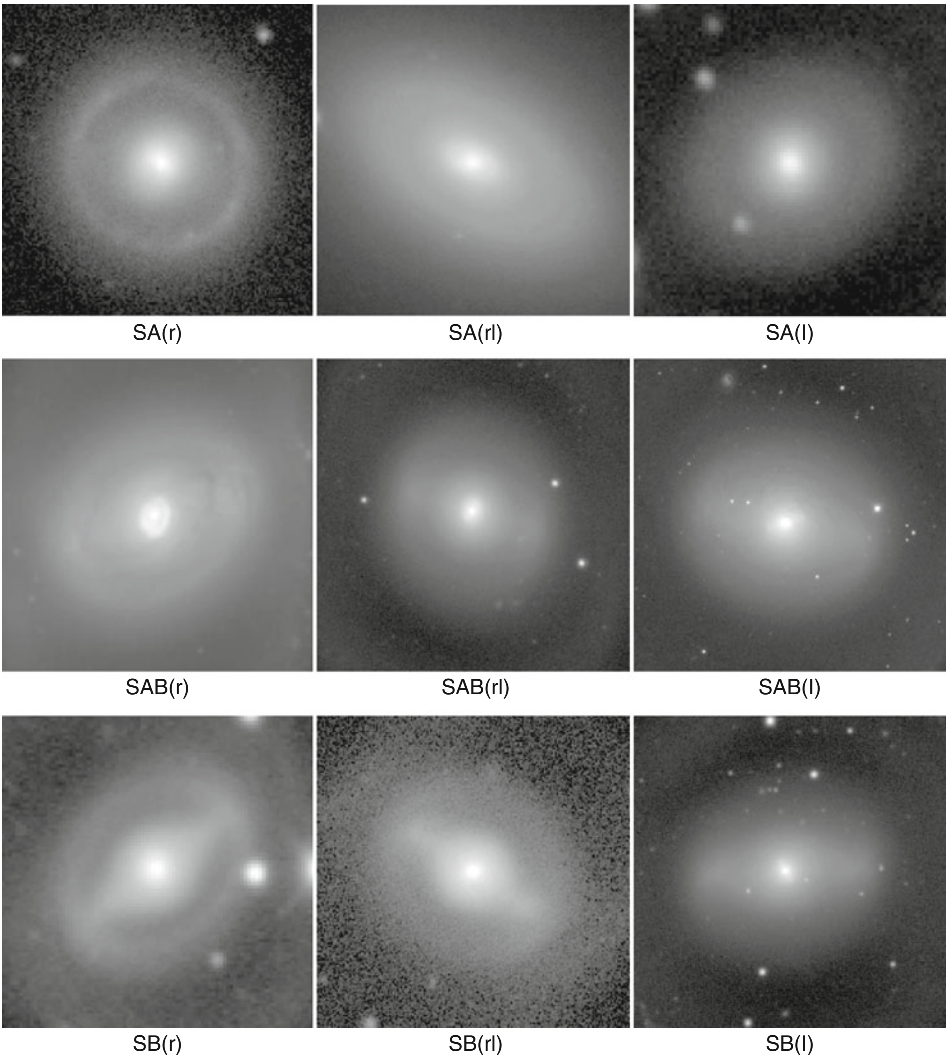
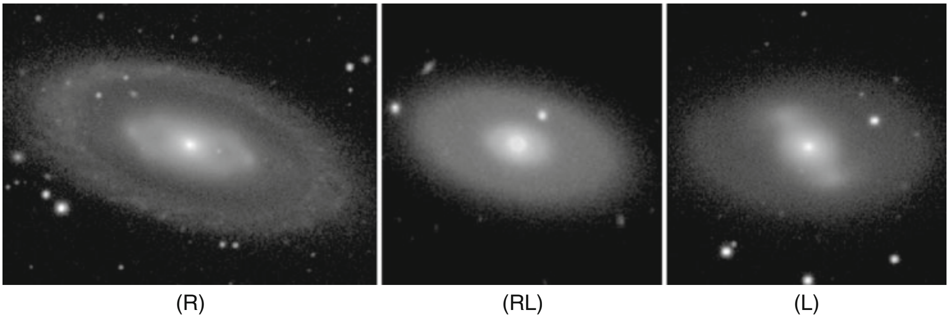
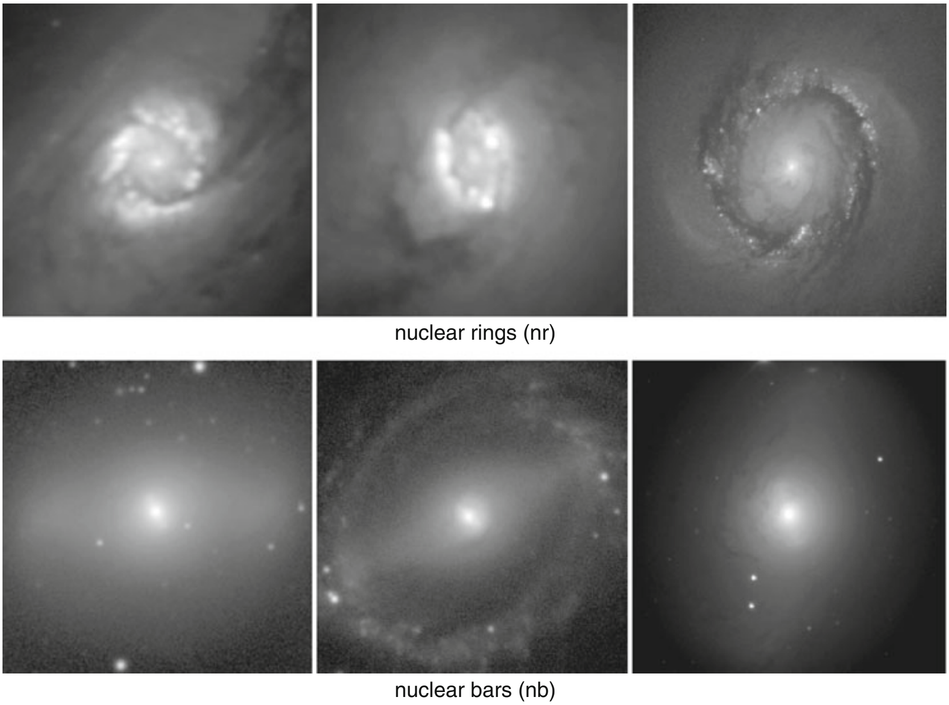


Fig. 1-16
Examples showing the continuity of inner rings (*r*) and lenses (*l*) for barred and nonbarred galaxies. The galaxies are (*left* to *right*): *Row 1* – NGC 7187, 1553, and 4909; *Row 2* – NGC 1326, 2859, and 1291; *Row 3* – ESO 426–2, NGC 1211, NGC 1543. All images are *B* or *g*-band from the dVA

Comerón et al. (2010) were also able to derive a reliable estimate of the relative frequency of nuclear rings as 20% ± 2% over the type range $S0^{-}$ to Sd, confirming with smaller error bars the previous result of Knapen (2005). Assuming that nuclear rings are a normal part of galaxy evolution, these authors argue that the rings may survive for 2–3 Gyr. Interestingly, it was also found that 19% ± 4% of nuclear rings occur in nonbarred galaxies, implying either that the rings may have formed when a bar was stronger (evidence of bar evolution) or that ovals or other mechanisms can lead to their formation. Mazzuca et al. (2009; see also Knapen 2010) connect some of the properties of nuclear rings to the rate at which the rotation curve rises in the inner regions.



■ **Fig. 1-17**
Examples showing the continuity of outer rings (*R*) and lenses (*L*). The galaxies are (*left* to *right*): NGC 7020 (dVA *B*), NGC 5602 (SDSS image), and 2983 (dVA *B*)


nuclear rings (nr)

nuclear bars (nb)

■ **Fig. 1-18**
Examples of nuclear rings and secondary (nuclear) bars. The galaxies are (*left* to *right*): *Row 1* – NGC 1097, 3351, and 4314; *Row 2* – NGC 1543, 5850, and 1291. All images are *B*-band from the dVA

The most extreme nuclear ring known is found in the SBa galaxy ESO 565–11 (see also ➜ Sect. 7). At 3.5 kpc in diameter, not only is the ring one of the largest known nuclear rings but also it has an extreme elongated shape compared to more typical nuclear rings.

Nuclear bars lie in the same radial zone as nuclear rings and sometimes lie inside a nuclear ring. Three examples are shown in the second row of ➜ *Fig. 1-18*. These average about one-tenth the size of a primary bar. There is no preferred angle between the axis of the nuclear bar and the primary bar, suggesting that the two features have different pattern speeds (Buta and Combes 1996; dVA).

Neither nuclear rings nor nuclear bars were recognized in the original Hubble–Sandage–de Vaucouleurs classifications, presumably in part because the use of small-scale photographic plates for extensive galaxy classification limited the detectability of the features in the (typically) overexposed centers. Modern multiband digital imaging greatly facilitates the detection of the small rings and bars, allowing their inclusion in the classification. Buta and Combes (1996) and Buta et al. (2010a) suggested the notation nr for nuclear rings and nb[4] for nuclear bars, respectively, to be used as part of the variety classification as in, for example, SB(r,nr)b for NGC 3351 or SAB(l,nb)0/a for NGC 1291. Continuity may exist for these features like other rings and primary bars. (For example, nuclear lenses (nl) may also be recognized.) In blue-light images, the appearance of the central region of a barred galaxy can be strongly affected by dust. For example, NGC 1365 shows a nuclear spiral in blue light, while in the infrared, the morphology is that of a nuclear ring (Buta et al. 2010a). The morphologies of some galaxies have a full complement of classifiable features. For example, accounting for all the rings and bars seen in NGC 3081, the classification is ($\mathrm{R}_1\mathrm{R}_2'$)SAB(r,nr,nb)0/a.

Lisker et al. (2006) use the terminology "S2B" for double-barred galaxies, a reasonable alternative approach to classifying these objects. Lisker et al. successfully identified nuclear bars in galaxies at redshifts $z = 0.10$–0.15 (from HST ACS observations), the most distant ones recognized thus far.

6.4 Spiral Arm Morphologies

A classification such as "Sb" tells one that a galaxy is a spiral of moderate pitch angle and degree of resolution of the arms and that a significant bulge may be present. The type does not directly tell (1) the multiplicity of the spiral pattern, (2) the character of the arms (massive, filamentary, grand design, or flocculent), or (3) the sense of winding of the arms (leading or trailing the direction of rotation). These are nevertheless additional dimensions to galaxy morphology.

The multiplicity of the spiral pattern refers to the actual number of spiral arms, usually denoted by the integer m. Examples of spirals having $m = 1$–5 are illustrated in ➲ *Fig. 1-19*. The multiplicity is not necessarily straightforward to determine and may be a function of radius. For example, a spiral may be two armed in the inner regions and multi-armed in the outer regions. Spirals of low m are usually *grand design*, a term referring to a well-defined global (meaning galaxy-wide) pattern of strong arms. The typical grand-design spiral has two main arms, as in NGC 5364 (lower left frame of ➲ *Fig. 1-19*). In contrast, a flocculent spiral has piecewise continuous arms but no coherent global pattern (Elmegreen 1981). NGC 5055 is an example shown in the middle left frame of ➲ *Fig. 1-19*. This category is relevant mainly to optical wavebands. In the infrared, an optically flocculent spiral like NGC 5055 reveals a more coherent global grand-design spiral (Thornley 1996; see also ➲ *Fig. 1-44*), indicating that dust is partly responsible for the flocculent appearance.

[4] In a study of galactic nuclei, van den Bergh (1995) proposed the notation "NB" for nuclear bars, although what he refers to are not the same as the features described here. However, the presence or absence of a nucleus is an important morphological issue that may be connected to evolutionary history, as noted by van den Bergh.

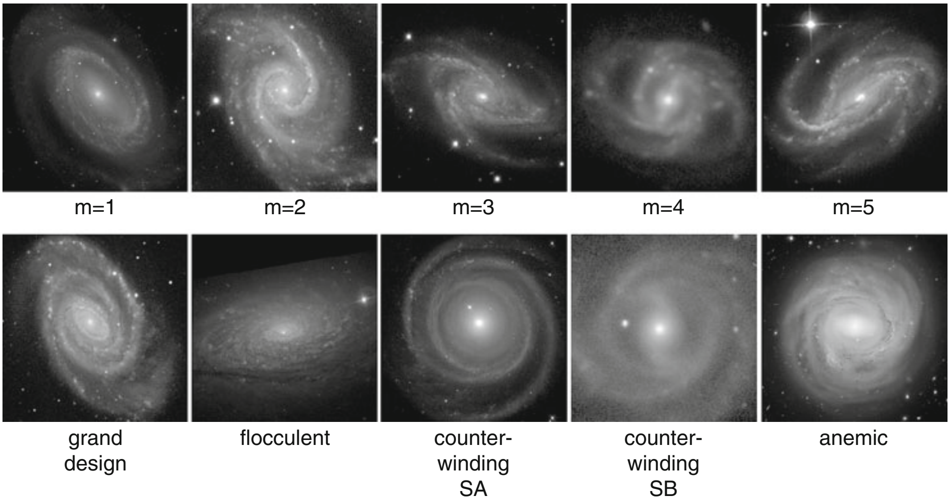


Fig. 1-19
Examples showing spiral arm character differences in the form of arm multiplicity, grand-design and flocculent spirals, counterwinding spirals, and an anemic spiral. The galaxies are (*left* to *right*): *Row 1* – NGC 4725, 1566, 5054, ESO 566–24, and NGC 613; *Row 2* – NGC 5364, 5055, 4622, 3124, and 4921. All images are *B*-band from the dVA except NGC 5055, which is SDSS *g*-band, and NGC 4921, which is from Hubble Heritage

The terms "massive" and "filamentary" arms are due to Reynolds (1927) and are discussed by Sandage (1961, 1975). Massive arms are broad, diffuse, and of relatively low contrast, as in M33, while filamentary arms are relatively thin in comparison and lined by knots or filaments, as in NGC 5457 (M101). De Vaucouleurs (1956) originally used these distinctions as part of his classification, but later dropped the references to spiral arm character probably because of the complexity it added to his types.

Elmegreen and Elmegreen (1987) used a different approach to spiral arm character by recognizing a series of spiral *arm classes* based on arm continuity and length (but not necessarily contrast). Ten classes ranging from flocculent (ACs 1–4) to grand design (5–12; numbers 10 and 11 were later dropped) are illustrated in *Fig. 1-20* (see Elmegreen and Elmegreen 1987 for a description of each class). Thus, spiral character is a well-developed additional dimension to galaxy classification. A simpler approach advocated by Elmegreen and Elmegreen is "G" for grand design, "F" for flocculent, and "M" for multiple armed. The arms of grand-design spirals are in general thought to be density waves and may in fact represent quasi-steady wave modes (e.g., Bertin et al. 1989; Zhang 1996, 1998, 1999), although there is also some evidence that spirals may be transient (see review by Sellwood 2010). Flocculent spirals may be sheared self-propagating star formation regions (Seiden and Gerola 1982).

Figure 1-19 also shows two examples of a new class of spirals, called *counterwinding* spirals. In these cases, an inner spiral pattern winds outward in the opposite sense to an outer spiral pattern. In the case of NGC 4622 (row 2, middle), the inner pattern has only a single arm and

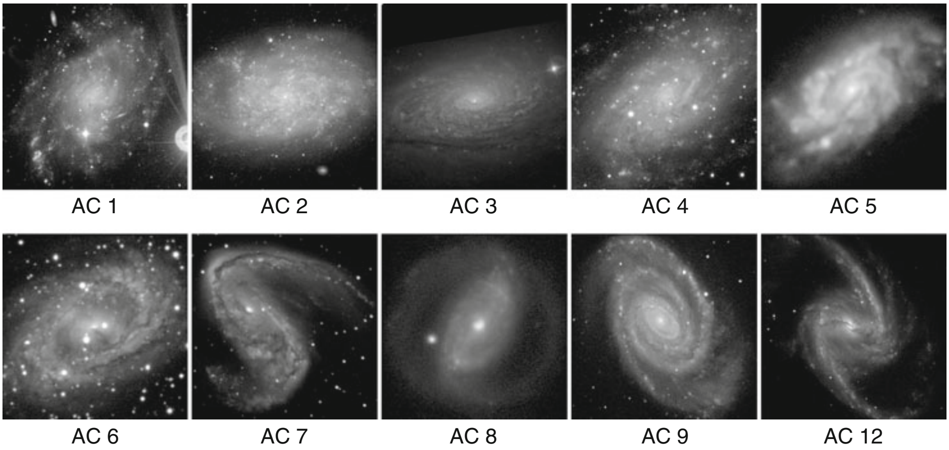


Fig. 1-20
Examples showing the spiral arm classes of Elmegreen and Elmegreen (1987). The galaxies are (*left* to *right*): *Row 1* – NGC 45, 7793, 5055, 2403, and 1084. *Row 2*: NGC 6300, 2442, 3504, 5364, and 1365. All images are *B*-band from the dVA, except NGC 5055 which is SDSS *g*-band

the outer pattern has two arms, while in NGC 3124 (row 2, middle right), the inner pattern is two armed and the outer pattern is at least four armed. The two cases are very different because NGC 4622 is essentially nonbarred while in NGC 3124, the inner spiral is classified as a bar. The presence of oppositely winding spiral patterns in the same galaxy means that one set of arms is trailing (opening opposite the direction of rotation) while the other set is leading (opening into the direction of rotation). In general, studies of the dust distribution as well as the rotation of spirals have shown that trailing arms are the rule (de Vaucouleurs 1958). Surprisingly, straightforward analysis of a velocity field and the dust pattern in NGC 4622 led Buta et al. (2003) to conclude that the strong outer two-armed pattern in this galaxy is leading, while the inner single arm is trailing. This led to the characterization of NGC 4622 as a "backward spiral galaxy," apparently rotating the wrong way. An additional nonbarred counterwinding spiral has been identified in ESO 297−27 by Grouchy et al. (2008). In this case, the same kind of analysis showed that an inner single arm leads while a three-armed outer pattern trails. No comparable analysis has yet been made for NGC 3124.

Väisänen et al. (2008) have shown that a two-armed (but not counterwinding) spiral in the strongly interacting galaxy IRAS 18293-3413 is leading. Even with this, the number of known leading spirals is very small (dVA). Leading spirals are not expected to be as long-lived as trailing spirals since they do not transfer angular momentum outward, and this is needed for the long-term maintenance of a spiral wave (Lynden-Bell and Kalnajs 1972).

An interesting example of leading "armlets" was described by Knapen et al. (1995a), who used K-band imaging of the center of the grand-design spiral M100 to reveal details of its nuclear bar and well-known nuclear ring/spiral. The nuclear bar has a leading twist that connects it to two bright K-band "knots" of star formation. This morphology was interpreted in terms of the expectations of gas orbits in the vicinity of an inner inner Lindblad resonance (IILR; Knapen et al. 1995b).

The final galaxy in ➋ *Fig. 1-19* is NGC 4921, an example of an *anemic* spiral. This is a type of spiral that is deficient in neutral atomic hydrogen gas, and, as a consequence, it has a lower amount of dust and star formation activity. The arms of NGC 4921 resemble those of an Sb or Sbc galaxy in pitch angle and extensiveness but are as smooth as those typically seen in Sa galaxies. Anemic spirals were first recognized as galaxies with "fuzzy" arms (see van den Bergh, 1998) where star formation has been suppressed due to ram-pressure stripping in the cluster environment. In the case of NGC 4921, the environment is the Coma Cluster. The idea is that such galaxies will eventually turn into S0 galaxies (van den Bergh 2009a). Anemic spiral galaxies are further discussed in ➋ Sect. 10.2.

Seigar et al. (2008) have demonstrated the existence of a correlation between spiral arm pitch angles and supermassive central black hole masses. The sense of the correlation is such that black hole mass is highest for the most tightly wound spirals and lowest for the most open spirals. The correlation is expected because black hole mass is tightly correlated to bulge mass and central mass concentration, and spiral arm pitch angle is tied to shear in galactic disks, which itself depends on mass concentration (Seigar et al. 2005).

6.5 Luminosity Effects

Luminosity effects are evident in the morphology of galaxies through surface brightness differences between giants and dwarfs and through the sophistication of structure such as spiral arms. Van den Bergh (1998) describes his classification system which takes luminosity effects into account using a set of luminosity classes that are analogous to those used for stars. The largest, most massive spirals have long and well-developed arms, while less massive spirals have less well-defined arms.

The nomenclature for the classes parallels that for stars: I (supergiant galaxies), II (bright giant galaxies), III (giant galaxies), IV (subgiant galaxies) and V (dwarf galaxies). Intermediate cases I–II, II–III, III–IV, and IV–V, are also recognized.

➋ *Figure 1-21* shows galaxies which van den Bergh (1998) considers primary luminosity standards of his classification system. The original van den Bergh standards for these classes were based on the small-scale paper prints of the Palomar Sky Survey. Sandage and Tammann (1981) adopted the precepts of the van den Bergh classes but revised the standards based on large-scale plates. In general, luminosity class I galaxies have the longest, most well-developed arms; luminosity class III galaxies have short, patchy arms extending from the main body; while luminosity class V galaxies have very low surface brightness and only a hint of spiral structure. The classes are separated by type in ➋ *Fig. 1-21* because among Sb galaxies, few are of luminosity class III or fainter, while among Sc and later-type galaxies, the full range of luminosity classes is found. van den Bergh does not use types like Sd or Sm for conventional de Vaucouleurs late-types but instead uses S^- and S^+ to denote "early" (smoother) and "late" (more patchy) subgiant spirals. Similarly, van den Bergh uses Sb^- and Sb^+ to denote "early" and "late" Sb spirals, respectively. (Some of these would be classified as Sab and Sbc by de Vaucouleurs.) According to the standards listed by van den Bergh (1998), an Sb I galaxy is 2–3 mag more luminous than an Sb III galaxy, while an Sc I galaxy is more than 4 mag more luminous than an S V galaxy.

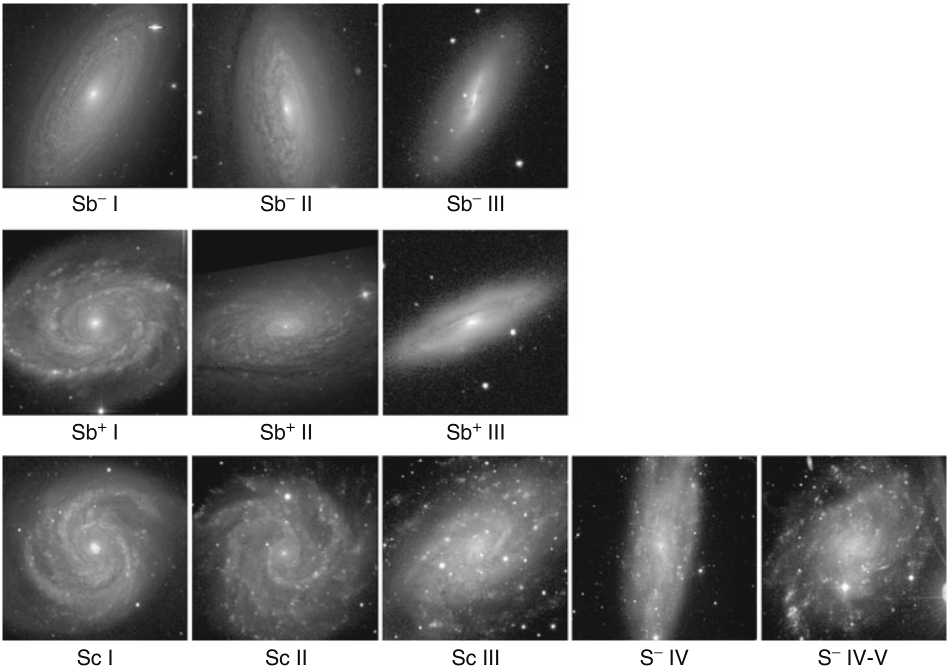


Fig. 1-21

Examples showing van den Bergh luminosity classes. The galaxies are (*left* to *right*): *Row 1* – NGC 2841 (dVA *B*), 3675 (SDSS *g*), and 4064 (SDSS *g*). *Row 2*: NGC 5371 (dVA *B*), 5055 (SDSS *g*), and 4586 (SDSS *g*). *Row 3*: NGC 4321, 3184, 2403, 247, and 45 (all dVA *B*). The classifications are in the van den Bergh (1998) system

7 The Morphology of Galactic Bars and Ovals

Bars are among the most common morphological features of disk-shaped galaxies. Unlike spiral arms, bars cross the “spiral-S0 divide” in the Hubble sequence and are abundant among spirals (at the 50–70% level) when both SAB and SB types are considered (de Vaucouleurs 1963; Sellwood and Wilkinson 1993). The bar fraction has cosmological significance (Sheth et al. 2008), and many estimates of the nearby galaxy bar fraction have been made from both optical and IR studies (see Buta et al. 2010a for a summary of recent work). Comerón et al. (2010) have argued that mass-limited samples provide the most reliable assessment of the structural evolution of bars over intermediate redshifts and showed that intermediate mass galaxies ($10^{10.5} < M/M_\odot < 10^{11}$) continue their bar evolution to $z \approx 0.2$, while high-mass disk galaxies ($M/M_\odot > 10^{11}$) reached a mature state (constant bar fraction with redshift) by $z \approx 1$.

Bars are fairly well-understood features of galaxy morphology that have been tied to a natural instability in a rotationally supported stellar disk (see review by Sellwood and Wilkinson 1993). The long-term maintenance of a bar in a mostly isolated galaxy is thought to depend on how effectively it transfers angular momentum to other galaxy components, such as the halo (Athanassoula 2003). Bars are thought to be transient features that, in spiral galaxies, may dissolve and regenerate several times over a Hubble time (Bournaud and Combes 2002). Alternatively, bars may be long-lived density wave modes that drive secular evolution of both the stellar and gaseous distributions (Zhang and Buta 2007; Buta and Zhang 2009). The possible secular evolution of bars in S0 galaxies is discussed by Buta et al. (2010b). Bars are also thought to drive spiral density waves (Kormendy and Norman 1979; Buta et al. 2009; Salo et al. 2010).

The actual morphology of bars shows interesting variations that merit further study. The family classifications SAB and SB indicate some measure of bar strength but do not allude to the varied appearances of bars even among those only classified as SB. Regular bars, such as those illustrated in *Fig. 1-8*, are the conventional types that define the SB class. *Figure 1-22* shows "ansae"-type bars, referring to bars which have "handles" or bright enhancements at the ends. Martinez-Valpuesta et al. (2007) carried out a statistical study and found that ansae are present in ≈40% of early-type barred galaxies and are very rare for types later than stage Sb. Ansae are usually detectable in direct images, but their visibility can be enhanced using unsharp-masking (all the right frames for each galaxy in *Fig. 1-22*). Morphologically, ansae may be small round enhancements like those seen in NGC 5375 and 7020, but in some cases, ansae are approximately linear enhancements, giving the bar a parallelogram appearance as in NGC 7098, or

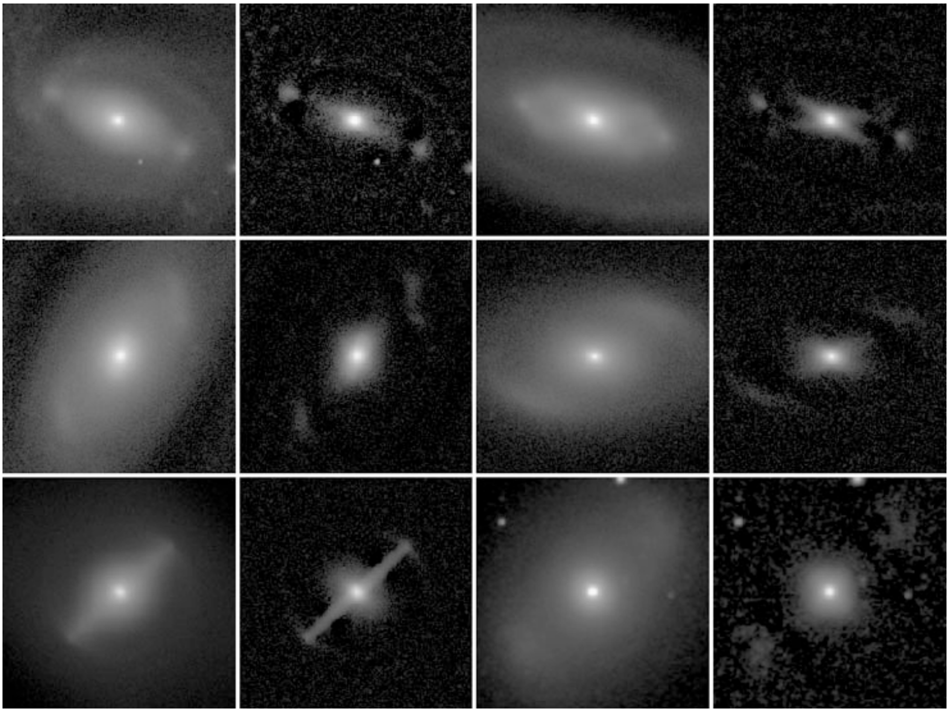


Fig. 1-22

Examples showing ansae bar morphologies as compared to one mostly non-ansae bar. For each galaxy, the *left* frame is the full image while the *right* frame is an unsharp-masked image, both in units of mag arcsec^{-2}. The galaxies are (*left* to *right*): *Row 1* – NGC 5375 (SDSS *g*) and 7020 (*I*) (both round ansae type); *Row 2* – NGC 7098 (*I*, linear, partly wavy ansae) and NGC 1079 (K_s, curved ansae); *Row 3* – NGC 4643 (*I*, mostly non-ansae type with trace of ring arcs at bar ends) and NGC 4151 (OSUBSGS *H*, irregular ansae)

curved arcs, giving the bar a partial ring appearance as in NGC 1079. Color index maps in the dVA show that ansae are generally as red as the rest of the bar, indicating the features are stellar dynamical in origin rather than gas dynamical. Nevertheless, ansae made of star-forming regions are known. Martinez-Valpuesta et al. (2007) illustrate the case of NGC 4151, a well-known Seyfert 1 galaxy with a strong bar-like inner oval. The appearance of this galaxy's ansae in the 1.65 μm H-band is shown in the lower right frames of ➋ *Fig. 1-22*, where the ansae are seen to have irregular shapes compared to the others shown.

Another subtlety about bars is their general boxy character. Athanassoula et al. (1990) showed that generalized ellipses fit the projected isophotes of bars better than do normal ellipses. For 11 or 12 SB0 galaxies examined in this study, a significant degree of boxiness was found near the bar semimajor axis radius.

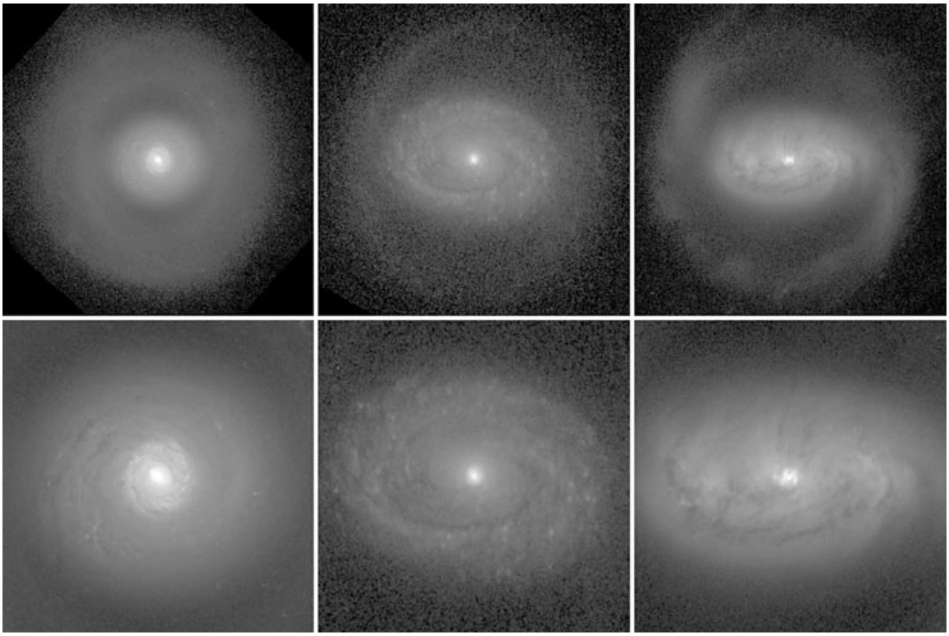
The unsharp-masked image of NGC 7020 in ➋ *Fig. 1-22* shows an X-shaped pattern in the inner regions that is the likely signature of a significantly three-dimensional bar. NGC 1079 and 5375 show hints of similar structure. The X-pattern is expected to be especially evident in edge-on galaxies which show the extended vertical structure of bars. Many examples have been analyzed (Bureau et al. 2006; see also the dVA). Buta et al. (2010a) have suggested that edge-on bars recognized from the X-pattern be denoted "SB$_x$."

The cause of bar ansae is uncertain. In simulations, Martinez-Valpuesta et al. (2006) found that ansae form late, after a second bar-buckling episode in a disk model with a live axisymmetric halo, and appear as density enhancements in both the face-on and edge-on views.

➋ *Figure 1-23* shows three examples of galaxies having oval inner disks. These features are described in detail by Kormendy and Kennicutt (2004) [KK04], who present both photometric and kinematic criteria for recognizing them. The images in ➋ *Fig. 1-23* are optical and have been cleaned of foreground and background objects and have also been deprojected based on the mean shape and major axis position angle of faint outer isophotes. In all three cases, the presence of an outer ring allows clear recognition of the oval shape, assuming that the inner and outer structures are in the same plane. The upper panels of ➋ *Fig. 1-23* show the full morphology with the outer rings, while the lower panels show the bright oval inner disks. The shapes of the oval disks are varied and range from axis ratio 0.84 for NGC 4736 to 0.55 for NGC 1808. The most striking example is NGC 4941, whose oval disk includes a bright, normal-looking spiral pattern with isophotal axis ratio 0.68. Many other examples are provided by KK04.

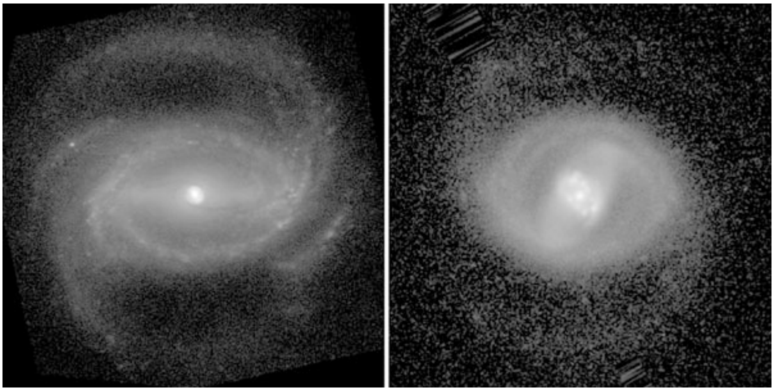
The ovals appear to play a bar-like role in these galaxies. The outer rings may be resonant responses to the nonaxisymmetric potential of the ovals, which clearly harbor a great deal of mass in spite of the mildness of their departures from axisymmetry. As noted by KK04, oval disk galaxies are expected to evolve secularly in much the same manner as typical barred galaxies. On the other hand, ovals themselves could be products of bar secular evolution. Laurikainen et al. (2009) found that the near-IR bar fraction in S0 galaxies is significantly less than that in S0/a or early-type spiral galaxies (also found by Aguerri et al. 2009 in the optical), while the oval/lens fraction is higher, suggesting that some ovals/lenses might be dissolved bars. Further evidence that bars might be dissolving in some galaxies is the detection of extremely weak bars in residual images of visually nonbarred S0 galaxies where a two-dimensional decomposition model has been subtracted. Such a bar is detected in the SA0° galaxy NGC 3998 (Laurikainen et al. 2009). Aguerri et al. (2009) suggest that central concentration is a key factor in bar evolution and that a unimodal distribution of bar strengths argues against the idea that bars dissolve and reform (Bournaud and Combes 2002).

Regular barred galaxies also often include an oval bounded by an elongated inner ring. The deprojected blue-light images of two examples are shown in ➋ *Fig. 1-24*. NGC 1433 has a very strong normal bar and one of the most intrinsically elongated inner rings known. The inner ring



◘ Fig. 1-23

Examples of oval disk galaxies. *Left* to *right*: NGC 4736 (*B*), NGC 4941 (*B*), and NGC 1808 (*V*). The images are cleaned of foreground and background objects and have been deprojected and rotated so that the major axis of the oval is horizontal. The *upper panels* show the ovals imbedded within outer rings, while the *lower panels* focus on the ovals alone



◘ Fig. 1-24

Examples of two galaxies having highly a elongated inner ring at the boundary of a broad oval. *Left* to *right*: NGC 1433, ESO 565−11 (both *B*-band). Each galaxy also has a prominent bar which is aligned with the oval and inner ring in NGC 1433, but misaligned with these features in ESO 565−11. ESO 565−11 also has a highly elongated, large nuclear ring of star-forming regions

lies at the edge of an oval which is more conspicuous at longer wavelengths. In NGC 1433, the inner ring, the oval, and the bar are all aligned parallel to each other. The situation is different in ESO 565−11, whose bright oval is strongly misaligned with a prominent bar, but similar to NGC 1433, the oval is bounded by an inner pseudoring. The suggestion in this case is that the bar and the oval are independent patterns.

8 Dust Morphologies

Dust lanes are often the most prominent part of the interstellar medium detectable in an optical image of a galaxy. ➲ *Figure 1-25* shows different classes of dust lanes using direct optical images on the left for each galaxy and a color index map on the right. The color index maps are coded such that blue features are dark while red features (like dust lanes) are light.

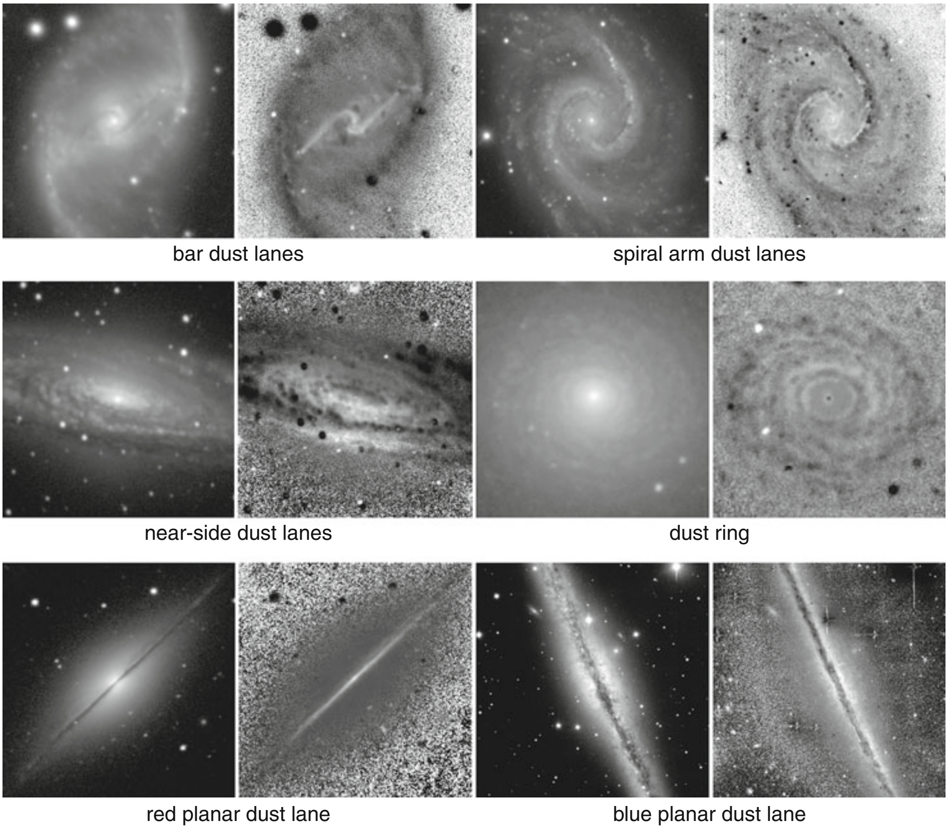


Fig. 1-25
Examples showing different classes of dust lanes (*left* to *right*) – *Row 1* NGC 1530, *V*-band image and $V - K_s$ color index map; NGC 1566, *B*-band and $B - K_s$ color index map; *Row 2* – NGC 7331, *B*-band image and $B - I$ color index map; NGC 7217, *B*-band image and $B - I$ color index map; *Row 3* – NGC 7814, *B*-band image and $B - I$ color index map; NGC 891, *B*-band and $B - V$ color index map

The bars of intermediate (mainly Sab to Sbc) spirals often show *leading* dust lanes, that is, well-defined lanes that lie on the leading edges of the bars, assuming the spiral arms trail. The example shown in ➲ *Fig. 1-25*, NGC 1530, has an exceptionally strong bar, and the lanes are very straight, regular, and well defined. These dust lanes are a dynamical effect associated with the bar. The lanes may be curved or straight. Athanassoula (1992) derived models of bar dust lanes and tied the curvature to the strength of the bar in the sense that models with stronger bars developed straighter dust lanes. Comerón et al. (2009) recently tested this idea by measuring the curvature of actual dust lanes as well as quantitative values of the bar strengths for 55 galaxies. They found that strong bars can only have straight dust lanes, while weaker bars can have straight or curved lanes.

In the same manner as bars, a strong spiral often has dust lanes on the concave sides of the inner arms. This is shown for NGC 1566 in the upper right panels of ➲ *Fig. 1-25*. Both bar and spiral dust lanes are face-on patterns. Another type of face-on pattern is the *dust ring*. The inner dust ring of NGC 7217 is shown in the right middle frames of ➲ *Fig. 1-25*, and it appears as the dark, inner edge of a stellar ring having the same shape. Dust rings can also be detected in more inclined galaxies.

An inclined galaxy with a significant bulge also can show another dust effect: in such a case, the bulge is viewed through the dust layer on the near side of the disk, while the dust is viewed through the bulge on the far side of the disk. This leads to a reddening and extinction asymmetry across the minor axis such that the near side of the minor axis is more reddened and extinguished than the far side. In conjunction with rotation data, this near-side/far-side asymmetry was used by Hubble (1943) and de Vaucouleurs (1958) to show that most spirals trail the direction of rotation.

The lower frames of ➲ *Fig. 1-25* show the planar dust lanes seen in edge-on spiral galaxies. The lane in NGC 7814 (lower left frames) is red which indicates that the galaxy is probably no later in type than Sa. This is consistent with the large bulge seen in the galaxy. However, the planar dust lane seen in NGC 891, type Sb, has a thin blue section in the middle of a wider red section. The blue color is likely due to outer star formation that suffers relatively low extinction. Some individual star-forming regions can be seen along the dust lane. In spite of the blue color, we are only seeing the outer edge of the disk in the plane.

Also related to galactic dust distributions are observations of *occulting galaxy pairs*, where a foreground spiral galaxy partly occults a background galaxy, ideally an elliptical (White and Keel 1992). With such pairs, one can estimate the optical depth of the foreground dust, often in areas where it might not be seen easily in an isolated spiral. An excellent example is described by Holwerda et al. (2009), who are able to trace the dust distribution in an occulting galaxy to 1.5 times than the standard isophotal radius.

Another way of illustrating the dust distribution in galaxies is with *Spitzer Space Telescope* Infrared Array Camera (IRAC) images at 8.0 μm wavelength. This is discussed further in ➲ Sect. 12.

9 The Morphologies of Galactic Bulges

A bulge is a very important component of a disk galaxy. In the context of structure formation in a cold dark matter (CDM) cosmology, bulges may form by hierarchical merging of disk galaxies, a process thought to lead to elliptical galaxies if the disks have approximately equal mass. Bulges formed in this way should, then, resemble elliptical galaxies, especially for early-type spirals. The bulges of later-type spirals, however, can be very different from the expectations of a merger-built bulge (also known as a "classical" bulge). In many cases, the bulge appears to be made of material associated with the disk.

KK04 reviewed the concept of "pseudobulges," referring to galaxy bulges that may have formed by slow secular movement of disk gas to the central regions (also known as disklike bulges; Athanassoula 2005). The main driving agent for movement of the gas is thought to be bars, which are widespread among spiral galaxies and which exert gravity torques that can move material by redistributing the angular momentum. Inside the corotation resonance, where the bar pattern speed equals the disk rotation rate, gas may be driven into the center to provide the raw material for building up a pseudobulge. KK04 review the evidence for such processes and argue that pseudobulges are a strong indication that secular evolution is an important process in disk-shaped galaxies.

➲ *Figure 1-26* shows the morphologies of both classical bulges and pseudobulges. The four classical bulge galaxies shown in the lower row, M31, NGC 2841, M81, and M104, have bright smooth centers and no evidence of bulge spiral structure or star formation. Classical bulges also tend to have rounder shapes than disks and can have significant bulge-to-total-luminosity ratios as illustrated by M104. Classical bulges are also more supported by random motions than by rotation. Many references to classical bulge studies are given by KK04. Formation mechanisms of such bulges are discussed in detail by Athanassoula (2005).

The two upper rows of ➲ *Fig. 1-26* are all pseudobulges as recognized by KK04. The first row shows HST wide V-band (filter F606W) images of the inner 1–1.3 kpc of four galaxies, NGC 3177, 4030, 5377, and 1353, in the type range Sa–Sbc. The areas shown account for much of the rise in surface brightness above the inward extrapolation of the outer disk light in these galaxies and would be considered bulges just on this basis. The HST images show, however, considerable spiral structure, dust, small rings, and likely star formation in these regions, characteristics not expected for a classical bulge. KK04 argue that instead these are pseudobulges that are highly flattened, have a projected shape similar to the outer disk light, have approximately exponential brightness profiles (Sersic index $n \approx 1$–2), and have a high ratio of ordered rotation to random motions. KK04 argue that a low Sersic index compared to $n = 4$ appears to be the hallmark of these pseudobulges, and a signature of secular evolution.

The second row in ➲ *Fig. 1-26* shows other kinds of pseudobulges discussed by KK04. NGC 6782 and 3081 (two left frames) have secondary bars, and KK04 considered that such features indicate the presence of a pseudobulge because bars are always disk features. In each case, the secondary bar lies inside a nuclear ring.

The other two galaxies in the second row of ➲ *Fig. 1-26*, NGC 128 and 1381, are examples of boxy or box-peanut bulges. These features have been linked to the vertical heating of bars, and if this is what they actually are, then KK04 argue that boxy and box-peanut bulges are also examples of pseudobulges. However, boxy and box/peanut bulges would *not* necessarily be the result of slow movement of gas by bar torques and subsequent star formation in the central regions but instead would be related to the orbital structure of the bar itself (Athanassoula 2005).

Recent studies have shown that pseudobulges are the dominant type of central component in disk galaxies. Although originally thought to be important only for late-type galaxies, Laurikainen et al. (2007) showed that pseudobulges are found throughout the Hubble sequence, including among S0-S0/a galaxies, based on sophisticated two-dimensional photometric decompositions. Such galaxies frequently have nuclear bars, nuclear disks, or nuclear

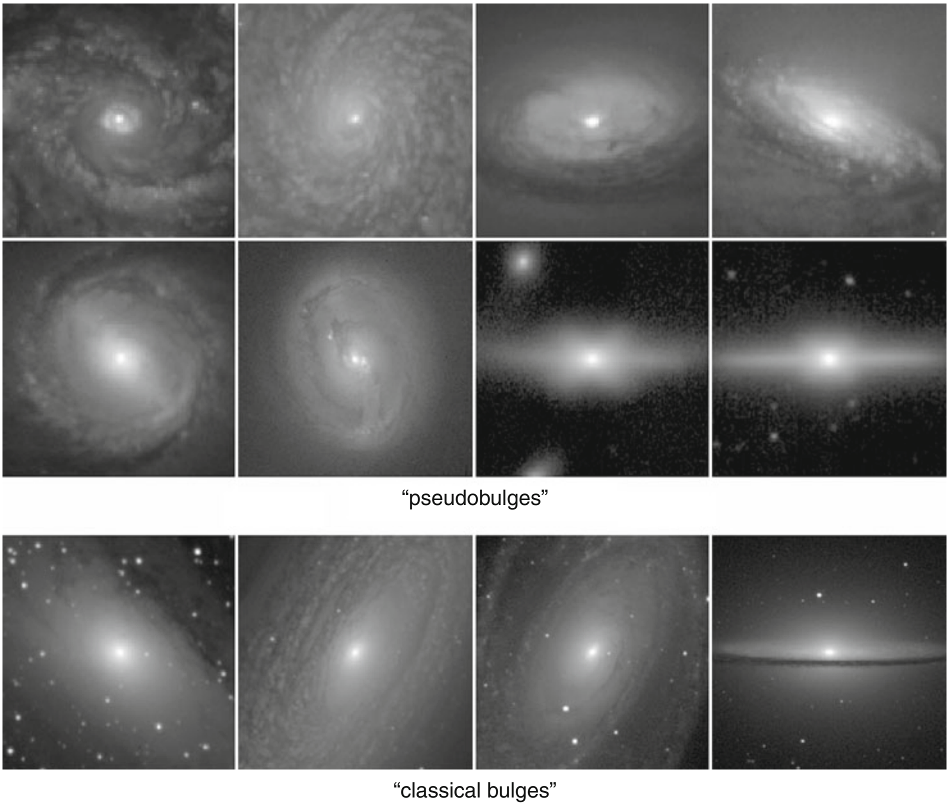


Fig. 1-26

Examples of pseudobulges and classical bulges in spiral galaxies (*left* to *right*): *Row 1*: NGC 3177, 4030, 5377, and 1353 (all HST wide *V*-band filter F606W; KK04); *Row 2*: NGC 6782 (*I*-band, F814W), 3081 (wide *B*, F450W), 128 (K_s), and 1381 (K_s); *Row 3*: NGC 224 (M31), 2841, 3031 (M81), and 4594 (M104) (all *B*-band). The images of NGC 128 and 1381 are from Bureau et al. (2006)

rings. Laurikainen et al. also found that bulge-to-total-flux (B/T) ratios are much less than indicated by earlier studies, especially for early Hubble types, and that the Sersic index averages $\lesssim 2$ across all types. The lack of gas in S0 and S0/a galaxies complicates the interpretation of their pseudobulges in terms of bar-driven gas flow and subsequent star formation. Instead, Laurikainen et al. link the pseudobulges in early-type galaxies to the evolution of bars. Laurikainen et al. (2010) also showed that S0s can have pseudobulges if they are stripped spirals, without invoking any bar-induced evolution.

Kormendy et al. (2010) showed that a significant fraction of large late-type disk galaxies in noncluster environments have no trace of a classical, merger-driven bulge. Some galaxies, like M33, may not even have a pseudobulge. Kormendy et al. note that the existence of so many massive pure disk galaxies challenges cold dark matter cosmology; they state: "In field environments, the problem of forming giant pure-disk galaxies in a hierarchically clustering universe is acute."

10 Effects of Interactions and Mergers

Galaxy morphology is replete with evidence for gravitational interactions, ranging from minor, distant encounters to violent collisions and major/minor mergers. Many of the most puzzling and exotic morphologies can be explained by interactions, and even sublimely normal galaxies, like ordinary ellipticals, have been connected to catastrophic encounters. Up to 4% of bright nearby galaxies are involved in a major interaction (Knapen and James 2009). In clusters, other types of interactions may occur, such as gas stripping and truncation of the star-forming disk. In this section, a variety of the types of morphologies that may be considered the results of external interactions are described and illustrated.

10.1 Normal Versus Catastrophic Rings

The three types of rings described so far, nuclear, inner, and outer rings, are aspects of the morphology of relatively normal, undisturbed galaxies. Inner rings and pseudorings are found in more than 50% of normal disk galaxies (Buta and Combes 1996), while outer rings and pseudorings are found at the 10% level. The latter rings could be more frequent because their faintness may cause them to go undetected, which is less likely to occur for inner rings. As has been noted, nuclear rings are found at the 20% level (Comerón et al. 2010). The high abundance of these types of ring features suggests that they are mainly products of *internal* dynamics, and in fact all three ring types have been interpreted in terms of internal processes in barred galaxies. The main interpretation of these kinds of rings has been in terms of orbital resonances with the pattern speed of a bar, oval, or spiral density wave. Resonances are special places where a bar can secularly gather gas owing to the properties of periodic orbits (Buta and Combes 1996). Buta (1995) showed that the intrinsic shapes and relative bar orientations of inner and outer rings and pseudorings support the resonance interpretation of the features. Schwarz (1981) suggested the outer Lindblad resonance for outer rings and pseudorings, while Schwarz (1984) suggested the inner 4:1 ultraharmonic resonance for inner rings and pseudorings and the inner Lindblad resonance for nuclear rings. Knapen et al. (1995b) and Buta and Combes (1996) provide further insight into these interpretations.

The resonance idea may only be valid in the case of weak perturbations. In the presence of a strong perturbation, the concept of a specific *resonance radius* can break down, although the idea of a broad *resonance region* could still hold (Contopoulos 1996). Regan and Teuben (2003, 2004) argue that nuclear rings and inner rings are better interpreted in terms of orbit transitions, that is, regions where periodic orbits transition from one major orbit family to another, as in the transition from the perpendicularly aligned x_2 family to the bar-aligned x_1 family (Contopoulos and Grosbol 1989).

Normal rings have also been interpreted in terms of "invariant manifolds" which emanate from the unstable L_1 and L_2 Lagrangian points in the bar potential (Romero-Gómez et al. 2006, 2007; Athanassoula et al. 2009a, b). This approach has also had some success in predicting the shapes and orientations of inner and outer rings, such as the R_1, R_1', $\mathrm{R}_1\mathrm{R}_2'$, and R_2' morphologies shown in ➲ *Figs. 1-14* and ➲ *1-15*. A morphology called "rR_1," which includes an oval inner ring and a figure eight-shaped R_1 ring (see NGC 1326 in ➲ *Fig. 1-15*), is especially well represented by this kind of model. The manifolds are tubes which guide orbits escaping the L_1 and L_2 regions. Note that in this interpretation, outer rings are not necessarily associated with the OLR (Romero-Gómez et al. 2006).

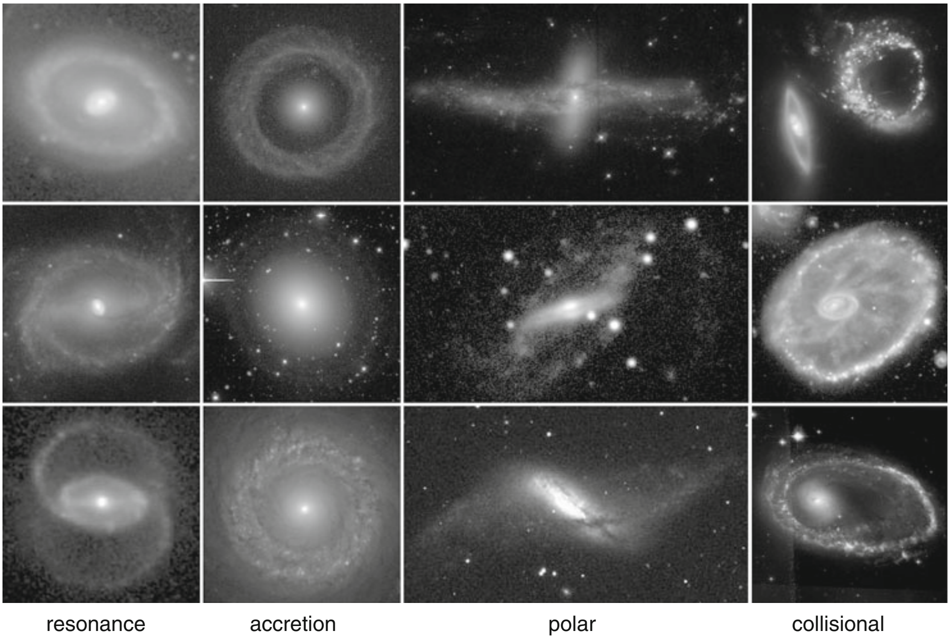


Fig. 1-27
Different classes of ring phenomena seen in galaxies (*top* to *bottom*): *Column 1* – NGC 3081, NGC 1433, and UGC 12646. *Column 2* – Hoag's Object, IC 2006, and NGC 7742; *Column 3* – NGC 4650A, ESO 235–58, and NGC 660; *Column 4* – Arp147, the Cartwheel, and the Lindsay-Shapley ring. All images are from the dVA except Arp 147, which is Hubble Heritage

Although the vast majority of the ringlike patterns seen in galaxies are probably of the resonance/orbit-transition/invariant-manifold type, other classes of rings are known that are likely the result of more catastrophic processes, such as galaxy collisions. *Figure 1-27* shows resonance rings in comparison to three other types: accretion rings, polar rings, and collisional rings (the latter commonly referred to as "ring galaxies"). The three accretion rings shown in Hoag's Object (Schweizer et al. 1987), IC 2006 (Schweizer et al. 1989), and NGC 7742 (de Zeeuw et al. 2002) are thought to be made of material from an accreted satellite galaxy. For IC 2006 and NGC 7742, the evidence for this is found in *counterrotation*: the material in the rings counterrotates with respect to the material in the rest of the galaxies. In Hoag's Object and IC 2006, the accreting galaxy is a normal E system.

Polar rings are also accreted features except that the accreting galaxy is usually a disk-shaped system, most often an S0 (Whitmore et al. 1990). In these cases, the accreted material comes in at a high angle to the plane of the disk. The configuration is most stable if the accretion angle is close to 90° or over the poles of the disk system. This limits the ability of differential precession to cause the ring material to quickly settle into the main disk. The polar feature can be a ring or simply an inclined and extended disk. Whitmore et al. (1990) present an extensive catalogue of probable and possible polar ring galaxies.

The main example illustrated in *Fig. 1-27* is NGC 4650A, where the inner disk component is an S0. Galaxies like NGC 4650A have generated considerable research because polar rings probe the shape of the dark halo potential (e.g., Sackett et al. 1994). The galaxies are also special because the merging objects have retained their distinct identities when most mergers lead to a single object. Brook et al. (2008) link the misaligned disks of polar ring galaxies to the process of hierarchical structure formation in a cold dark matter scenario.

While polar rings are most easily recognizable when both disks are nearly edge-on to us, cases where one or the other disk is nearly face-on have also been recognized. One example, ESO 235−58 (Buta and Crocker 1993) is shown in the middle-right panel of ➋ *Fig. 1-27*. In this case, the inner component is almost exactly edge-on and shows a planar dust lane and is likely a spiral rather than an S0. The ring component is inclined significantly to the plane of this inner disk but may not be polar. The faint outer arms in this component caused ESO 235−58 to be misclassified as a late-type barred spiral in RC3. Spiral structure in polar disks has been shown to be excitable by the potential of the inner disk, which acts something like a bar (Theis et al. 2006).

An example where the main disk is seen nearly face-on is NGC 2655 (Sparke et al. 2008). In this case, the polar ring material is seen as silhouetted dust lanes at an uncharacteristic angle to the inner isophotes. NGC 2655 also shows evidence of faint shells/ripples, indicative of a recent merger (➋ Sect. 10.3.3). Sil'chenko and Afanasiev (2004) have discussed NGC 2655 and other similar examples of inner polar rings in terms of the triaxiality of the potential.

Also illustrated in ➋ *Fig. 1-27* is NGC 660, which was listed as a possible polar ring galaxy by Whitmore et al. (1990). Like ESO 235−58, NGC 660 has an aligned dust lane in its inner disk component, which thus is likely to be a spiral, not an S0. The extraplanar disk is actually far from polar, being inclined only 55° (van Driel et al. 1995). A recent study of massive stars in the ring is given by Karataeva et al. (2004).

Collisional ring galaxies (Arp 1966; Appleton and Struck-Marcell 1996) are thought to be cases where a larger galaxy suffers a head-on collision with a smaller galaxy down its polar axis. The collision causes an expanding density wave of massive star formation, and multiple rings are possible. Three examples are shown in ➋ *Fig. 1-27*. Theys and Spiegel (1976) have discussed various classes of ring galaxies. Arp 147 (Arp 1966) is an example of type "RE," referring to a sharp elliptical ring with an empty interior. The Cartwheel (Higdon 1995) and the Lindsay-Shapley ring (Arp and Madore 1987) are examples of type "RN," meaning an elliptical ring with an off-center nucleus. Not shown in ➋ *Fig. 1-27* is a third category called "RK," where a single, large knot lies on one side of the ring, making the system very asymmetric.

Madore et al. (2009) have published a comprehensive atlas of all known likely collisional ring galaxies, many taken from the catalogue of Arp and Madore (1987). Based on this study, only 1 in 1,000 galaxies is a collisional ring galaxy. For entry, the Madore et al. atlas requires at least two objects in the immediate vicinity of the ring that might plausibly be the intruder galaxy. Most of the rings are not in cluster environments, however. The atlas also brings attention to several double-ring collisional systems, which have been predicted by numerical simulations (see Struck 2010 for a review). The unusual radial "spokes" in the Cartwheel, a feature not seen in any other collisional ring galaxy, could be related to interactive accretion streams (Struck et al. 1996).

Romano et al. (2008) present images of several ring galaxies that show the precollision stellar disk. They also show that rings are generally delineated by blue knots and that the off-centered nuclei are usually more yellow in color. In addition, some of the companion galaxies show diffuse asymmetric outer light, suggesting that they are being stripped.

➋ *Figure 1-27* shows that accretion rings can account for some of the rings seen in nonbarred galaxies. Buta and Combes (1996) argue that a bar is an essential element in resonance ring formation. ESO 235−58 shows that a polar ring-related system can resemble a ringed, barred galaxy. The three collisional rings are all very distinctive from the others.

10.2 Environmental Effects on Star-Forming Disks

Galaxy clusters are excellent laboratories for detecting the effects of environment on galaxy morphology and structure. Frequent mergers and environmental conversion of spirals into S0s are thought to be at the heart of the morphology-density relation, where early-type galaxies dominate cluster cores, and spirals and irregulars are found mainly in the outer regions (Dressler 1980; van der Wel et al. 2010). More detailed discussion of environmental effects on galaxies is given in Chap. 3 of this Volume.

The issue of environmental effects has a direct bearing on how we might interpret the Hubble sequence. For example, the continuity of galaxy morphology certainly seems apparent from the discussions in previous sections of this review. The Hubble sequence E-S0-Sa-Sb-Sc-I appears physically significant when total colors, mean surface brightnesses, and HI mass-to-blue-light ratios are considered, and the way features are recognized in the classification systems also favors the continuity. Morphological continuity does *not*, however, automatically imply that the galaxy types are in fact ordered correctly. For example, although Hubble placed S0s as a transition type between elliptical galaxies and spirals, this placement has been questioned by van den Bergh (1998, 2009a). Based on a statistical analysis of types given in the RSA, van den Bergh showed that S0 galaxies are typically 0.8–1.0 mag less luminous than E and Sa galaxies, implying that S0 galaxies, on the whole, cannot really be considered intermediate between E and Sa galaxies.[5] The preponderance of S0 galaxies in clusters led to the early suggestion (e.g., Spitzer and Baade 1951; Gunn and Gott 1972; Moore et al. 1996) that some type of external environmental interaction was responsible for stripping a spiral galaxy of its interstellar medium. If this actually occurred, then, as suggested by van den Bergh (2009a), this could imply that S0 galaxies have lost a substantial fraction of their spiral mass due to interactions. Alternatively, van den Bergh (2009b) argues that stripping of a lower luminosity, late-type spiral should be easier than stripping of a higher luminosity, early-type spiral, which could also account for the luminosity difference. In an examination of the environment of S0 galaxies, van den Bergh (2009b) found no significant difference in the average luminosities, flattenings, or distribution of S0 subtypes in clusters, groups, or the field, indicating that some S0s develop as a result of internal effects, such as the influence of an active galactic nucleus.

Barway et al. (2011) noted that lower luminosity S0s have a higher bar fraction than higher luminosity S0s (21% vs. 6%), suggesting that the two groups form in different ways (see also Barway et al. 2007). These authors suggest that faint S0s are stripped late-type spirals, which are known to have a high bar fraction (Barazza et al. 2008). However, Barway et al. applied no inclination restriction on their sample.

Environmental effects in clusters do not always have to involve drastic transformations in morphology. Sometimes the effects are more subtle. *Figure 1-28* shows several spiral galaxies that are also members of the Virgo Cluster. These galaxies highlight processes that affect the star-forming disk while leaving the older stellar disk relatively unaffected. NGC 4580 and 4689

[5] In contrast to van den Bergh's study of RSA S0 galaxies, Laurikainen et al. (2010) found that the absolute K_s-band magnitudes of a well-defined sample of S0s are similar to those of early-type spirals in the OSUBSGS sample. The sample was mostly drawn from RC3 and includes some galaxies classified as ellipticals in RC3 and as S0s in the RSA (see Sect. 12).

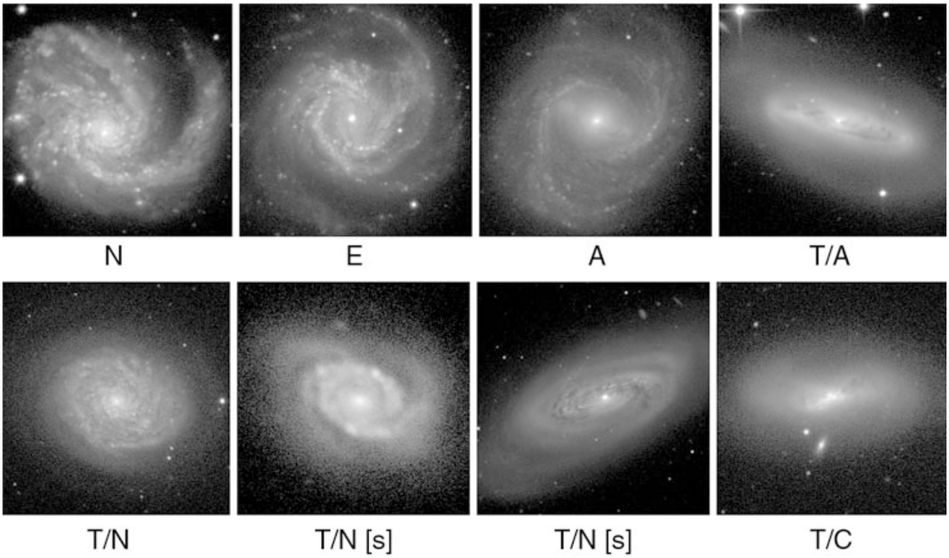


Fig. 1-28
Blue-light morphologies of eight Virgo Cluster spirals having different Koopmann and Kenney (2004) Hα star formation morphologies. The galaxies are (*left* to *right*): *Row 1*: NGC 4254 (normal N); NGC 4303 (enhanced *E*); NGC 4548 (anemic *A*); NGC 4293 (truncated/anemic *T/A*); *Row 2*: NGC 4689 (truncated/normal *T/N*); NGC 4580 and 4569 (truncated/normal (severe) *T/N* [s]); NGC 4424 (truncated/compact *T/C*). Images are dVA *B*, except for NGC 4424 which is SDSS *g*

are galaxies having a patchy inner disk and a smooth outer disk, called "Virgo types" by van den Bergh et al. (1990). These objects suggest that the environment of such galaxies has somehow truncated the star-forming disk, with a greater concentration of truncated disks toward the cluster core. Similar results are obtained from observations of the HI gas disks of Virgo Cluster galaxies (e.g., Giovanelli and Haynes 1985; Cayatte et al. 1994; Chung et al. 2009).

Koopmann and Kenney (2004) summarize the results of an extensive survey of Hα emission from Virgo Cluster galaxies and identify different categories of environmentally influenced star formation characteristics based on Hα imaging. The blue-light images of examples of each category are included in *Fig. 1-28* and show how the subtleties are manifested in regular morphology. Using a sample of isolated spiral galaxies to define "normal" star formation, Koopmann and Kenney defined several categories of Virgo spiral galaxy star-forming disks: Category "N" refers to disks whose star formation is within a factor of 3 of the normal levels. "E" cases have star formation enhanced by more than a factor of 3 compared to normal. "A" cases are "anemic" spirals (Sect. 6.4) having star formation reduced by more than a factor of 3 compared to normal. "T/N" refers to galaxies where the star-forming disk is sharply cut off, but inside the cutoff, the star formation levels are normal (the [s] means truncation is severe). One of these, NGC 4580, is so unusual that Sandage and Bedke (1994) classify it as Sc(s)/Sa, where the Sc part is the inner disk and the Sa part is the outer disk. In "T/A galaxies," the inner star formation is at a low level, as in anemic cases, while in "T/C" galaxies, most of the star formation is confined within the inner 1 kpc. Koopmann and Kenney found that the majority of Virgo Cluster spiral galaxies have truncated star-forming disks.

The idea is that the interstellar medium (ISM) of a cluster galaxy can interact with the intracluster medium (ICM), stripping the ISM (via ram pressure; Gunn and Gott 1972) but leaving the stellar disk intact. Truncated gas and star-forming disks result because ram-pressure stripping is more severe in the outer parts of galaxies (e.g., Book and Benson 2010 and references therein). In Virgo, most galaxies with truncated star-forming disks have relatively undisturbed stellar disks and normal to slightly enhanced inner disk star formation rates, suggesting that ICM-ISM stripping is the main mechanism in the reduction of their star formation rates. The cases found to have relatively normal or enhanced star formation rates are preferentially located in the outer parts of the cluster and likely have never visited the core region. Only galaxies which go near the center get significantly stripped. However, tidal effects also contribute to morphological changes. Several galaxies, including many of the T/C class, display peculiarities consistent with tidal effects, such as nonaxisymmetric circumnuclear star formation, shell features (e.g., NGC 4424 in ➲ *Fig. 1-28*), and enhanced inner star formation rates.

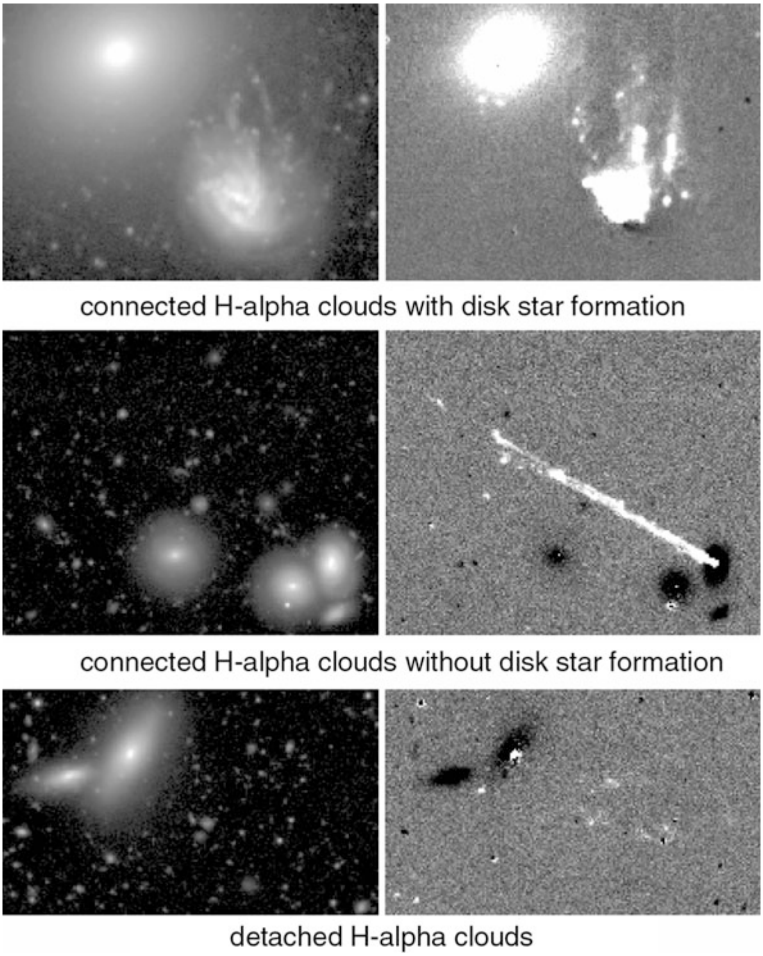
A recent study by Yagi et al. (2010) provides dramatic and clear evidence of disk gas stripping in galaxies thought to be relatively new arrivals to the core region of the Coma Cluster. Using deep Hα imaging, these authors detected ionized gas in clouds that are mostly outside the main disk of a dozen Coma galaxies. Three distinct morphologies of the distributions of these clouds were found: (1) connected clouds that blend with disk star formation; (2) long, connected lines of clouds that extend from a central gas knot but are not related to the disk light; and (3) clouds completely detached from the main disk. Examples of these categories are illustrated in ➲ *Fig. 1-29*. Yagi et al. interpret them in terms of an evolutionary gas-stripping sequence where category (1) galaxies are in an earlier phase of stripping while the category (3) galaxies are in the most advanced phase. It is likely that large disk galaxies in Coma would be completely stripped eventually because of the cluster's high ICM density and broad velocity distribution. The same process seen in Coma likely occurs in Virgo but is only partial for the large spirals owing to the lower ICM density and velocities in Virgo (Koopmann and Kenney 2004).

10.3 Interacting and Peculiar Galaxies

10.3.1 Tidal Tails, Arms, Bridges, and Streams

It is perhaps fitting that the first major spiral galaxy discovered was in the interacting pair M51 (➲ Sect. 2). Numerical simulations (Salo and Laurikainen 2000a, b) have shown that both parabolic and bound passages of the companion, NGC 5195, can explain the observed morphology and other characteristics of the system. It turns out that M51 defines a class of interacting systems known as M51-type pairs. ➲ *Figure 1-30* shows an example in the pair NGC 2535-6. In each case, the larger component has a strong two-armed spiral, with one arm appearing "drawn" to the smaller companion. An extensive catalogue of M51-type pairs is provided by Jokimaki et al. (2008).

Other distant encounters can produce tidal tails or bridges of material between galaxies (top row, ➲ *Fig. 1-30*). NGC 4676, also known as the "Mice," is a pair of strongly interacting galaxies where a very extended tidal tail has formed in one component. The strongly interacting pair NGC 5216/18 has developed a bright connecting bridge of material, and each component shows tidal tails. The evolution of this system, and the role of encounters on bar formation, is described by Cullen et al. (2007).



connected H-alpha clouds with disk star formation

connected H-alpha clouds without disk star formation

detached H-alpha clouds

Fig. 1-29

Three galaxies in a possible evolutionary stripping sequence in the Coma Cluster. The images and categories are from Yagi et al. (2010). The *left* frames are *B*-band images in units of mag arcsec^{-2}, while the *right* frames are net Hα images in linear intensity units (called NB–R by Yagi et al.). From *top* to *bottom*, the galaxies are GMP 3816, GMP 2910, and GMP 2923. The idea is that GMP 3816 is in an earlier phase of stripping, such that there is still considerable disk ionized gas; GMP 2910 is in a more advanced phase with still connected clouds but an absence of disk emission; and finally GMP 2923 is in the most advanced phase of the three, showing only scattered HII regions

Recent extremely deep imaging of otherwise well-known nearby normal disk galaxies has revealed the presence of enormous tidal streams of faint light in the halos of the galaxies. One of the best-known cases, NGC 5907, was shown by Martínez-Delgado et al. (2008) to have a spectacular set of large, arcing loops at high angle to the disk plane, attributed to tidal disruption of a single small satellite galaxy in a near-circular orbit accretion event. The arcs are thought to be Gyr old and are analogous to the Galactic Sagittarius stream. A pilot survey of other galaxies (Martínez-Delgado et al. 2010) has revealed a variety of faint features attributable

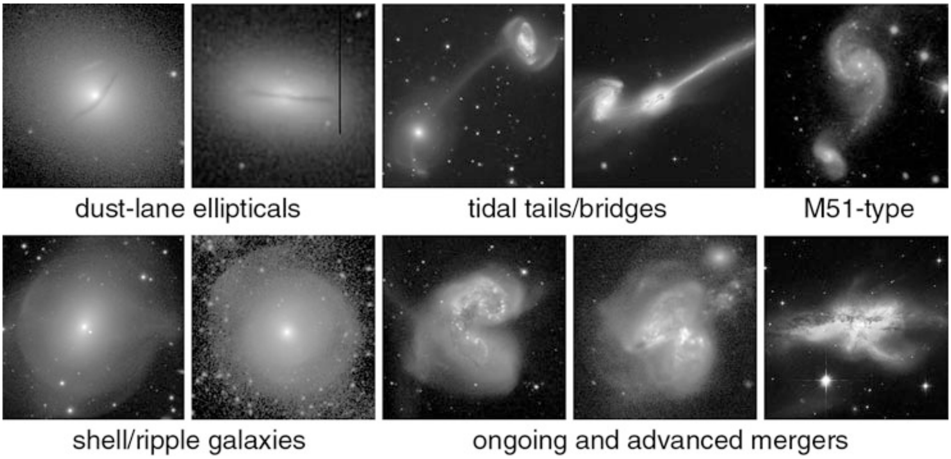


Fig. 1-30
Interacting and peculiar galaxies. The galaxies are (*left* to *right*): *Row 1*: NGC 5485 (SDSS), 4370 (SDSS), 5216/18 (copyright Adam Block/Mt. Lemmon SkyCenter/U. Arizona), 4676 (Hubble Heritage), and 2535-6 (SDSS); *Row 2*: NGC 2865, 474, 4038-9, 3690, and 6240, all *B* except NGC 474, which is a 3.6 μm image (Sect. 12)

to accretion events, including giant plumes, spikes, and umbrella-like structures, the latter categories attributed to low-mass satellites accreted in more radial orbits. Tidal streams can be thought of as evidence for hierarchical formation, whereby stellar halos are built up from the debris of small satellite mergers (Λ Cold Dark Matter cosmology; e.g., Steinmetz and Navarro 2002).

10.3.2 Dust-Lane Ellipticals

Figure 1-30 also shows several examples of morphologies that may result from minor mergers of a small galaxy with a more massive, pre-existing elliptical galaxy. Bertola (1987) brought attention to the unusual class of *dust-lane ellipticals*, where an otherwise normal elliptical galaxy shows peculiar lanes of obscuring dust. It was de Vaucouleurs's personal view that "if an elliptical shows dust, then it's not an elliptical!" However, Bertola showed that an unusual case like the radio elliptical galaxy NGC 5128, where a strong dust lane lies along the *minor axis* of the outer light distribution, is simply the nearest example of a distinct class of objects. Further study showed that dust-lane ellipticals come in several varieties. The minor axis dust-lane type appears most common, but cases of alignment along the major axis of the outer isophotes (major axis dust lanes) as well as cases of misalignment are also known (see the upper left panels of *Fig. 1-30*). The origin of these very regular dust lanes is thought to be a merger of a gas-rich companion (e.g., Oosterloo et al. 2002). The regularity of the dust lanes suggests that the mergers are in advanced states.

10.3.3 Shell/Ripple Galaxies

The two lower left frames of *Fig. 1-30* show examples of galaxies having "shells," or faint, arc-shaped brightness enhancements of varying morphology. They were first discovered on deep photographs by Malin (Malin and Carter 1980) and appeared to be associated mainly with elliptical galaxies. In fact, the first examples, NGC 1344 and 3923, are classified in catalogues as ordinary ellipticals because the shells are not detectable on photographs of average depth. Once the class was recognized, a detailed search led to other examples which were listed by Malin and Carter (1980, 1983). The term "shells" implies a particular three-dimensional geometry that Schweizer and Seitzer (1988) argued imposes a prejudice on the interpretation of the structures. They proposed instead the alternate term "ripples," which implies less of a restrictive geometry.

The explanation of shell/ripple galaxies is one of the great success stories in galactic dynamics (see review by Athanassoula and Bosma 1985). Shells are thought to be remnants of a minor merger between a massive elliptical and a lower mass disklike galaxy. The main requirements are that the disk-shaped galaxy be "cold," or lack any random motions, and that the potential of the elliptical galaxy should be rigid, meaning the elliptical is much more massive than its companion. The smaller galaxy's stars fall into the center of the galaxy and phase wrap or form alternating outward-moving density waves made of the disk galaxy's particles near the maximum excursions of their largely radial orbits in the rigid potential. Many, but not all, of the main properties of shell Es can be explained by this model. Other issues concerning shell galaxies are reviewed by Kormendy and Djorgovski (1989).

Taylor-Mager et al. (2007; see their Fig. 2) have proposed a simple classification of interacting systems that highlights different interaction classes. A premerger (type pM) includes two interacting galaxies that are sufficiently far apart to suffer little apparent distortion. A minor merger (mM) is two galaxies showing evidence of merging, but one component is much smaller than the other. A major merger (M) has two comparable brightness galaxies in the process of merging, while a merger remnant (MR) is a state sufficiently advanced that the merging components are no longer distinct. The three lower right frames in ➋ *Fig. 1-30* show examples of these types.

10.3.4 Ultraluminous Infrared Galaxies

Related to interacting systems are the infrared-bright galaxies first identified by Rieke and Low (1972) based on 10 μm photometry. From studies based on the Infrared Astronomical Satellite (IRAS), Sanders and Mirabel (1996) classified a galaxy as a "luminous infrared galaxy" (LIRG) if its luminosity in the 8–1,000 μm range is between 10^{11} and 10^{12} $L_\odot$. If the luminosity in the same wavelength range exceeds 10^{12} $L_\odot$, then the object is called an "ultraluminous infrared galaxy" (ULIRG). Detailed studies have shown that at high redshifts, LIRGS and ULIRGS are a dominant population of objects (see discussion in Pereira-Santaella et al. 2010).

The morphologies of nine ULIRGS were studied using HST B and I-band images by Surace et al. (1998). Their montage of six of these objects is shown in ➋ *Fig. 1-31*. In every case there are clear signs of interactions, and all are likely linked to mergers or mergers in progress. Several, like Mk 231, have bright Seyfert nuclei. Arribas et al. (2004) obtained extensive imaging of local LIRGS and found a similar high proportion of strongly interacting and merging systems.

The merger rate is considered one of the most important parameters for understanding galaxy evolution. It has been difficult to estimate, and issues connected with it are discussed by Jogee et al. (2009; see also Conselice 2009). A merger is considered major if it involves a companion ranging from one fourth to approximately equal mass to the main galaxy. A major merger of two spiral galaxies can destroy both disks and lead to an $r^{\frac{1}{4}}$ law-profile remnant through violent relaxation. Minor mergers involve companions having one tenth to one fourth the mass of the

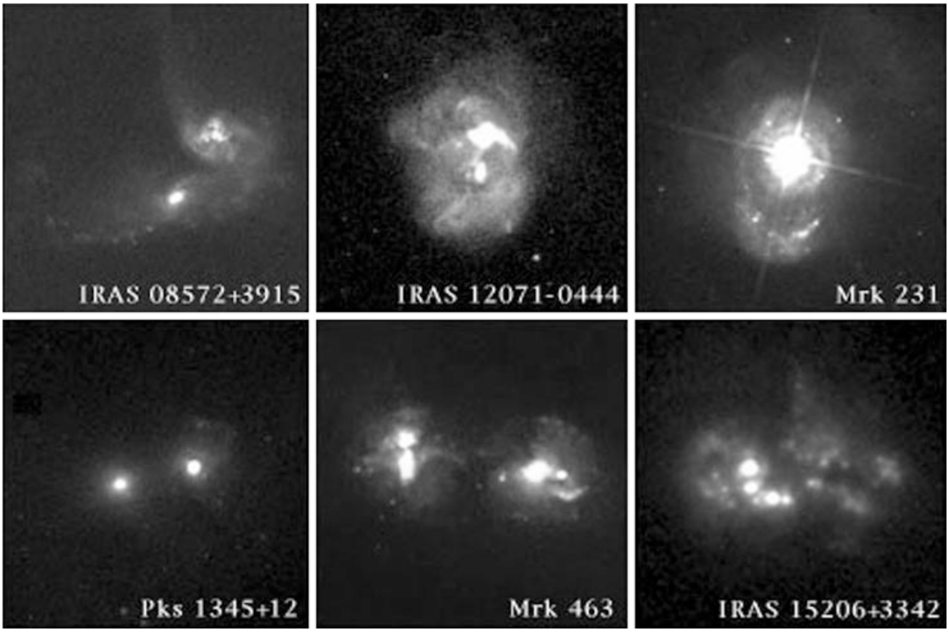


Fig. 1-31
The morphologies of six "ultraluminous infrared galaxies" from HST optical/near-IR imaging (Surace et al. 1998). These images are not in units of mag arcsec^{-2}

main galaxy. Both types of mergers, while in progress, can lead to many specific morphological features such as highly distorted shapes, tidal tails and bridges, shells and ripples, and warps. Even some bars and spiral patterns are thought to be connected to interactions, and especially galaxies with a double nucleus are thought to be mergers. As discussed in Sect. 10.1, mergers or collisions may also be at the heart of rare morphologies such as ring and polar ring galaxies. Using visual classifications of merger types, Jogee et al. (2009) estimate that 16% of high-mass galaxies have experienced a major merger, while 45% have experienced a minor merger during the past 3–7 Gyr ($z = 0.24$–0.80).

10.4 Warps

A warp is an apparent bend or slight twist in the shape of the disk of a spiral galaxy (see Sellwood 2010 and references therein). In a warp, stars and gas clouds move in roughly circular orbits, but the orientation of these orbits relative to the inner disk plane changes with increasing radius. Warps are most easily detected in edge-on galaxies because the bending of the outer orbits makes the galaxy look like an integral sign. Although often most pronounced in an HI map, warps can be seen in ordinary optical images of edge-on galaxies. *Figure 1-32* shows three galaxies having strong optical warping of the disk plane. In two of the galaxies, the bright inner disk is unwarped, while a fainter and thicker outer disk zone is twisted relative to the inner disk. In UGC 3697, the warping is exceptionally visible. In general, optical warping is less severe than HI warping.

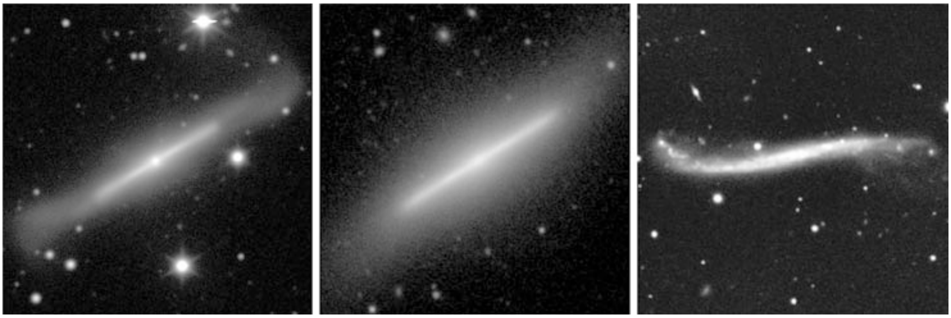


Fig. 1-32
Three galaxies showing strong optical disk warping. *Left* to *right*: NGC 4762 (*B*), NGC 4452 (SDSS), UGC 3697 (D. Darling, Internet Encyc. of Science)

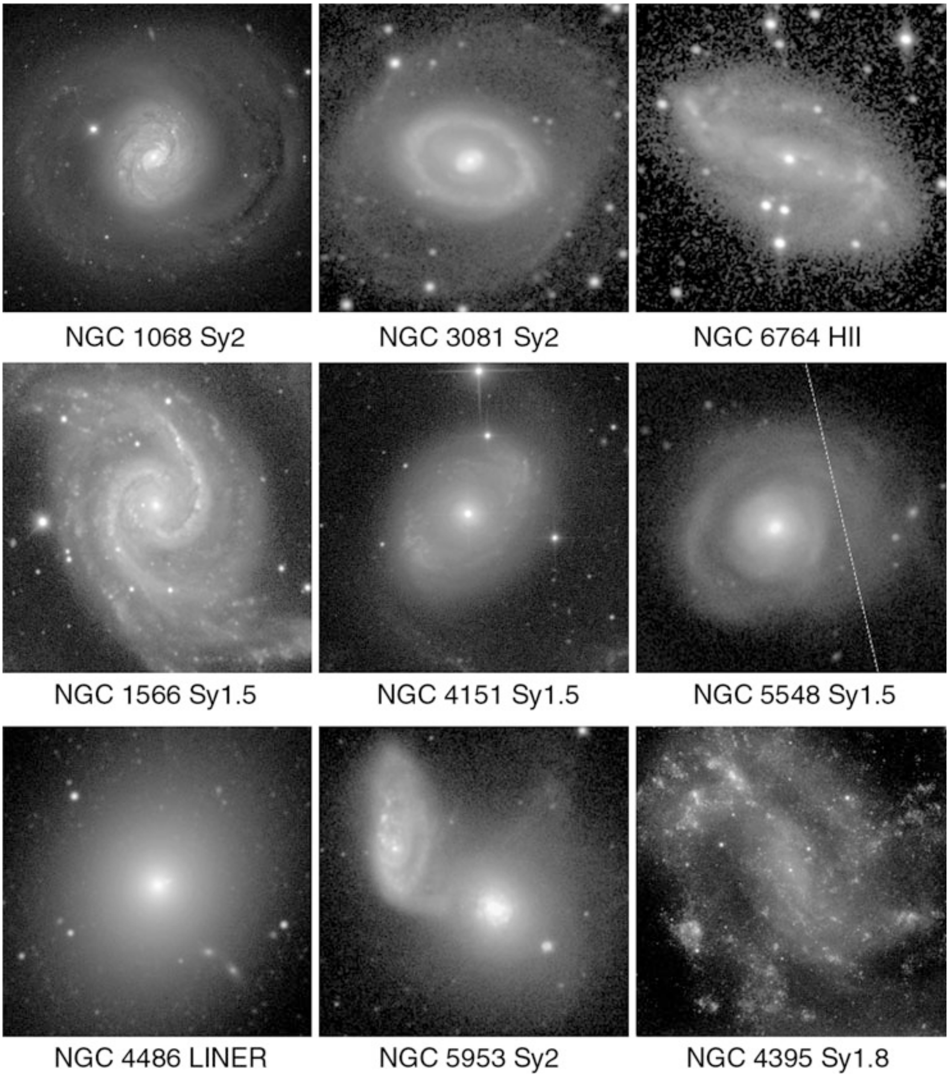
Warping is a very common aspect of spiral galaxies (e.g., Binney 1992) and has been interpreted in terms of perturbations (gaseous infall or interactions) that trigger bending instabilities (e.g., Revaz and Pfenniger 2007). Garcia-Ruiz et al. (2002) estimated HI warp angles, the angle between the inner disk plane and the assumed linear warping zone, to range from nearly 0° to more than 30°. A useful summary of previous warp studies is provided by Saha et al. (2009), who examine warp onset radii in mid-IR images. The theory of warps is reviewed by Sellwood (2010).

10.5 The Morphology of Active Galaxies

The morphology of active galaxies (also called "excited" galaxies by van den Bergh 1998) is important to consider because of a possible link between morphological features and the fueling of the active nucleus. Early studies showed a preponderance of ring, pseudoring, and bar features in Seyfert galaxies that suggested the link was bar-driven gas flow (Simkin et al. 1980). Several examples of the morphology of Seyfert and other active galaxies are shown in *Fig. 1-33*. The activity classifications are based mainly on spectroscopy, not on morphology, and are described by Veron-Cetty and Veron (2006).

A detailed study of active galaxy morphologies by Hunt and Malkan (1999) provided similar results to the early studies. These authors examined the morphologies of a large sample of galaxies selected on the basis of their 12 μm emission and found that outer rings and inner/outer ring combinations are three to four times higher in Seyfert galaxies than in normal spirals. In contrast, bars were found to occur with the same frequency (≈69%) in Seyferts as in normal spirals, while for HII/starburst galaxies, the frequency was much higher (>80%). Although outer rings are found mostly in barred galaxies, bars do not promote the nuclear activity of Seyfert galaxies. Hunt and Malkan (1999) interpret this inconsistency in terms of timescales: it takes roughly 3×10^9 years for a closed outer ring to form, a timescale during which a bar may weaken or dissolve. Because of this, a high ring frequency in Seyferts would indicate an advanced evolutionary state. Related to the same issue, Comerón et al. (2010) found that nuclear rings do not correlate with the presence of nuclear activity.

The study of Hunt and Malkan (1999) used mostly RC3 classifications to deduce the bar fraction in active galaxies. These visual classifications are based on blue-light images, and hence dust



NGC 1068 Sy2

NGC 3081 Sy2

NGC 6764 HII

NGC 1566 Sy1.5

NGC 4151 Sy1.5

NGC 5548 Sy1.5

NGC 4486 LINER

NGC 5953 Sy2

NGC 4395 Sy1.8

Fig. 1-33

Images of nearby active galaxies, dVA *B*-band except for NGC 5548 and 5953, which are SDSS. The activity classification is from Veron-Cetty and Veron (2006)

could effectively obscure some bars. Knapen et al. (2000) used high-resolution near-IR images of well-defined samples and quantitative bar detection methods to deduce that bars are more frequent in Seyfert galaxies than in a control sample of nonactive galaxies: 79% ± 7.5% versus 59% ± 9% (see also Laine et al. 2002). Laurikainen et al. (2004) came to a similar conclusion for 180 galaxies in the OSUBSGS, based on near-IR *H*-band images. The former studies used ellipse fits to identify bars, while Laurikainen et al. used Fourier analysis.

McKernan et al. (2010) also consider outer rings and pseudorings as probes of models of AGN fueling from interactions and mergers. The idea is that a closed outer ring takes a long time to form and is very fragile, being sensitive to interactions and changes in the bar pattern speed (e.g., Bagley et al. 2009). An interaction can change a closed outer ring into a pseudoring and could possibly destroy the ring. Thus, rings are probes of the interaction history of active galaxies. McKernan et al. found no difference between the AGN found in ringed galaxies and those found in galaxies without rings. But in those with rings, recent interactions can be ruled out and activity may be tied to short-term internal secular evolutionary processes.

Bahcall et al. (1997) presented HST images of 20 luminous, low redshift quasars observed with a wide V-band filter. ➲ *Figure 1-34* is reproduced from their paper and shows images of the host galaxies after removal of most of the quasar light. The images show a variety of morphologies, including ellipticals, interacting pairs, systems with obvious tidal disturbances, and normal-looking spirals. An example of the latter is PG1402 + 261 ($z = 0.164$), which is type (R_1')SB(rs)a based partly also on the image shown in Fig. 7 of Bahcall et al. From the number of

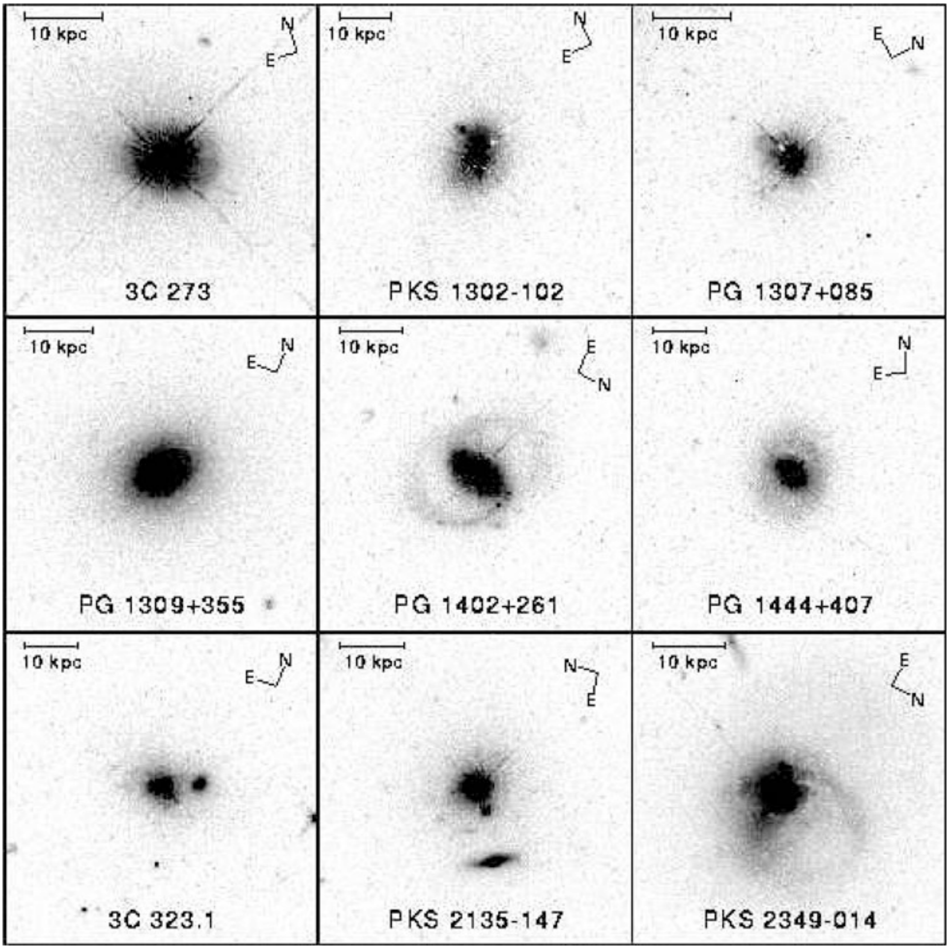


Fig. 1-34

The morphologies of the host galaxies of nine nearby quasars, from Bahcall et al. (1997). These images are not in units of mag arcsec^{-2}

hosts showing signs of interactions, as well as the number of companions, Bahcall et al. conclude that interactions may trigger the quasar phenomenon.

10.6 The Morphology of Brightest Cluster Members

Matthews et al. (1964) observed the optical morphologies of the radio sources identified with the brightest members of rich galaxy clusters. They found that the most common form was what Morgan (1958) called a "D" galaxy, meaning a galaxy having an elliptical-like inner region surrounded by an extensive envelope (see discussion in van den Bergh 1998). Although these superficially resembled Hubble's S0s, none were found having a highly elongated shape, implying that the galaxies are not as highly flattened as typical S0s. Another characteristic of the cluster D galaxies was their very large linear size and exceptional luminosity, much larger than a typical cluster member. To denote these extreme objects in the Morgan system, the prefix c was added as in the old classification of supergiant stars. Even today, Morgan's notation "cD" is used to describe these supergiant galaxies which are generally considered outside the scope of the Hubble system.

The most detailed study of the photometric properties of brightest cluster members (BCMs) was made by Schombert (1986, 1987, 1988; see also various references therein). The main BCM types Schombert considered were gE (giant ellipticals), D, and cD, distinguished mainly from the appearance of profile shape. D galaxies are larger and more diffuse with shallower profiles than gE galaxies, while a cD galaxy is the same as a D galaxy but with a large extended envelope (Schombert 1987). cD envelopes can extend to 500 kpc or more. Kormendy and Djorgovski (1989) argue that only cD galaxies are sufficiently physically distinct from ellipticals to merit being a separate class and recommended that the "D" class not be used.

Two cD galaxies and two gE galaxies are shown in ➲ *Fig. 1-35*. To give an idea of the scale of these objects, the vertical dimension of the frames corresponds to 201, 232, and 132 kpc for (left to right) UGC 10143 (A2152), NGC 4874/89 (A1656), and NGC 6041 (A2151), respectively. The cD classification of NGC 4874 is due to Schombert (1988), and one can see in ➲ *Fig. 1-35* that it is much larger and has a shallower brightness profile than nearby NGC 4889. The cD envelope

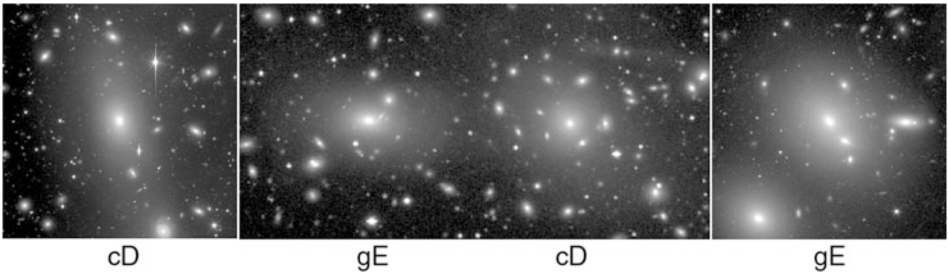


Fig. 1-35
Deep images of the brightest members of three rich galaxy clusters (*left* to *right*): UGC 10143 in A2152 (*R*-band), NGC 4889 (*left*) and 4874 (*right*) in A1656 (*B*-band), and NGC 6041 (*R*-band) in A2151. The images of UGC 10143 and NGC 6041 are from Blakeslee (1999), while that of NGC 4874-89 is from the dVA

is detected as an excess of light in the outer regions relative to a generalized brightness profile and may not even be the light that leads to the visual classification of cD.

Based on structural deviations such as the large radii, shallow profile slopes, and bright inner regions, Schombert (1987) concluded that BCMs fit well with the predictions of merger simulations, including accretion and "cannibalism" of smaller cluster members. Properties of cD envelopes (as separated photometrically from the parent galaxy) may suggest a stripping process for their formation (Schombert 1988).

As noted in ➋ Sect. 5.1, many BCMs in RC3 received the classification E^+, suggesting that the characteristic brightness profiles give a hint of an envelope interpreted as an incipient disk. The distribution of axial ratios of cDs actually is flatter on average than normal ellipticals (Schombert 1986), but it is not clear that the perceived envelopes in BCM E^+ galaxies are actually as flattened as a typical disk. A local example of a gE galaxy is M87, classified as type E^+0-1 by de Vaucouleurs.

11 Star Formation Morphologies

11.1 Hα Imaging

The standard waveband for galaxy classification, the *B*-band, is sensitive enough to the extreme population I component that the degree of resolution of spiral arms into star-forming complexes is part of the classification. The *B*-band, however, also includes a substantial contribution from the older stellar background. One way to isolate only the star-forming regions in a galaxy is imaging in Hα, which traces HII regions. Apart from showing the distribution of star formation (modified by extinction), Hα imaging also traces the rate of global photoionization, which in turn directly traces the rate of formation of stars more massive than about 10 $M_\odot$ (Kennicutt et al. 1994). The initial mass function (IMF), either assumed or constrained in some way (using, e.g., broadband colors), allows Hα fluxes to be converted to the total star formation rate over all stellar masses.

Hα imaging often shows well-organized patterns of HII regions that follow structures like spiral arms and especially rings. Images of six early-to-intermediate-type ringed galaxies are shown in ➋ *Fig. 1-36* (Crocker et al. 1996). The way Hα imaging usually works is a galaxy is imaged in or near its redshifted Hα wavelength and then in a nearby red continuum wavelength. The net Hα image is the difference between the Hα image and the scaled red continuum image. Often, the Hα filter used is broad enough to include emission from [NII] 6548 to 6584. For each of the galaxies shown in ➋ *Fig. 1-36*, the left image is the red continuum image, while the right image is the net Hα image. Of the four barred spirals shown (NGC 1433, 7329, 6782, and 7267), three show no HII regions associated with the bar. These three (NGC 1433, 7329, and 6782) all have conspicuous inner rings which are lined with HII regions. As shown by Crocker et al. (1996), the distribution of HII regions around inner rings is sensitive to the intrinsic shape of the ring. When the ring is highly elongated, the HII regions concentrate around the ends of the major axis (which coincide with the bar axis; NGC 6782), while when the ring is circular, the HII regions are more uniformly spread around the ring (NGC 7329). Inner ring shapes do not correlate well with maximum relative bar torques in a galaxy (dVA), but Grouchy et al. (2010) have shown that when local forcing is considered instead, ring shapes and bar strengths are well correlated.

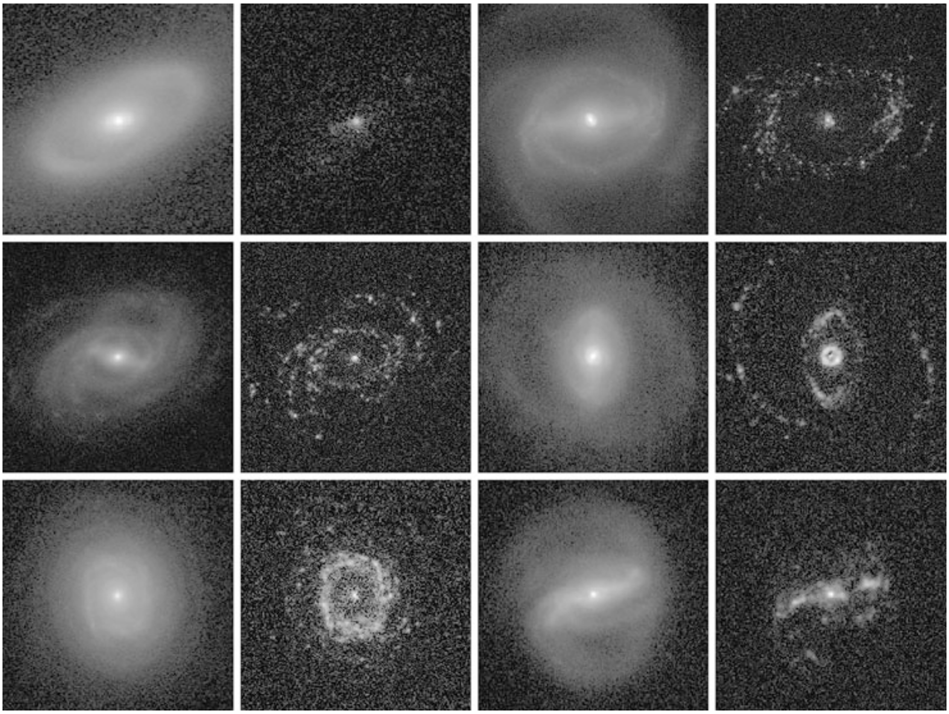


Fig. 1-36

Red continuum (*left*) and net Hα (*right*) images of six early-to-intermediate-type galaxies. The galaxies are (*left* to *right*), *top row*, NGC 7702, 1433; *middle row*, NGC 7329, 6782; *bottom row*, NGC 6935, 7267 (From Crocker et al. (1996))

The case of NGC 7267 is different in that most of its Hα emission is concentrated in the bar. Martin and Friedli (1997) have argued that star formation along galactic bars could provide clues to gas flow in the inner regions of galaxies and the fueling of starbursts and AGN. These authors present models which suggest an evolution from a pure Hα bar to an Hα bar with ionized gas in the center, to a gas-poor bar with strong nuclear or circumnuclear star formation. This suggests that the bar of NGC 7267 is younger than those in NGC 1433, 7329, and 6782.

The two other galaxies shown in *Fig. 1-36* are nonbarred or only weakly barred. NGC 6935 is an interesting case where a nonbarred galaxy has a strong ring of star formation. Grouchy et al. (2010) found that the star formation properties of inner rings, but not the distribution of HII regions, are independent of the ring shapes and bar strengths in a small sample. The case of NGC 7702 is different from the others. This is a late S0 (type S0^+) showing a very strong and largely stellar inner ring. The ring shows little ionized gas and appears to be in a quiescent phase of evolution.

11.2 Ultraviolet Imaging

The best global imaging of nearby galaxies at ultraviolet wavelengths has been obtained with the *Galaxy Evolution Explorer* (GALEX, Martin et al. 2005), which provided detailed images

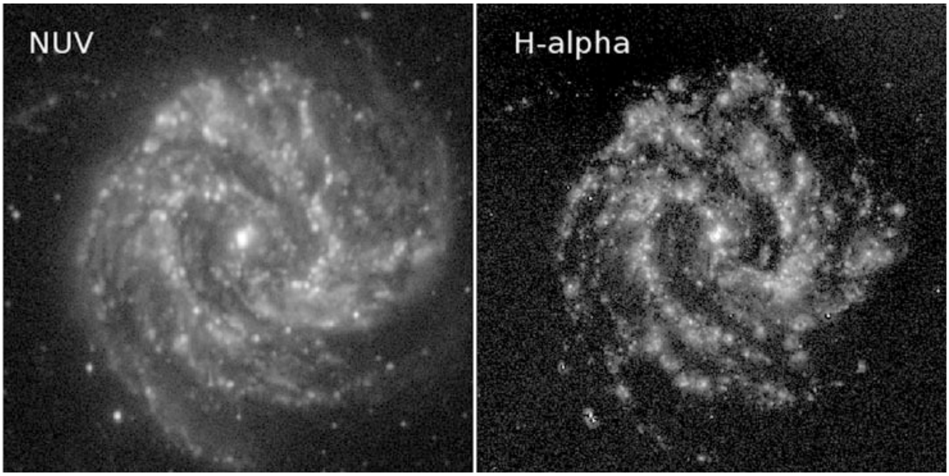


Fig. 1-37
Comparison of a near-UV image (0.225 μm) with an Hα image for the nearby spiral galaxy M83

of galaxies of all types at wavelengths of 0.225 and 0.152 μm. These images reveal young stars generally less than ≈100 Myr old but are affected by dust extinction. There is a strong correlation between the UV morphology and the Hα morphology (*Fig. 1-37*). Like Hα, UV fluxes from galaxies can be used to estimate star formation rates once extinction is estimated and are particularly sensitive to the ratio of the present to the average past star formation rate. UV imaging is an effective way of decoupling the recent star formation history of a galaxy from its overall, long-term star formation history.

Figure 1-38 shows near-UV (0.225 μm) images of eight galaxies over a range of Hubble types. The two earliest types, NGC 1317 and 4314 (both Sa), show a near-UV morphology dominated by a bright circumnuclear ring of star formation. The SB(r)b galaxy NGC 3351 shows a conspicuous inner ring of star formation and little emission from its bar region. In contrast, the SBc galaxy NGC 7479 shows strong near-UV emission from its bar. The late-type galaxies NGC 628 [type SA(s)c] and NGC 5474 [type SA(s)m] are typical of their types, but most interesting is NGC 4625. A key finding of GALEX was extended UV emission well beyond the optical extent of some galaxies. NGC 4625 is an example where the main optical part [type SABm] is only a small fraction of the extent of the UV disk (Gil de Paz et al. 2005).

The final object shown in *Fig. 1-38* is NGC 5253, a basic example of what de Vaucouleurs classified as I0 (Sect. 5.3). The inner region is a bright boxy zone of star formation, and even the extended disk is prominent.

11.3 Atomic and Molecular Gas Morphology

Related to star formation morphologies are the distributions of atomic and molecular gas. Far from being randomly distributed clouds of interstellar material, atomic and molecular gas morphologies can be highly organized, well-defined patterns. Recent high-quality surveys have

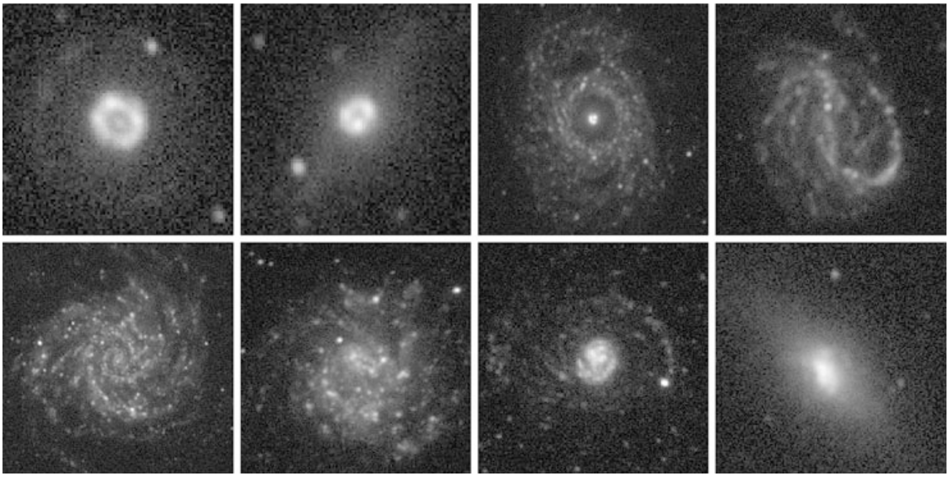


Fig. 1-38
GALEX near-UV images of eight galaxies: (*left* to *right*) *Top row*, NGC 1317, 4314, 3351, and 7479; *bottom row*, NGC 628, 5474, 4625, and 5253

provided some of the best maps of these distributions in normal galaxies. Atomic hydrogen is mapped with the 21-cm fine structure emission line, which has the advantage of not being affected by extinction and for being sufficiently optically thin in general to allow total HI masses to be derived directly from HI surface brightness maps. In addition, HI maps provide information on the kinematics and dynamics of the ISM, as well on the existence and distribution of dark matter in galaxies. Molecular hydrogen is generally mapped using the ^{12}CO J = 1 – 0 rotational transition at a wavelength of 2.6 mm, under the assumption that CO and hydrogen mix in a roughly fixed proportion.

Although numerous maps have been made of the HI distribution in nearby galaxies, the most sophisticated and detailed survey made to date is the "The HI Nearby Galaxies Survey" (THINGS, Walter et al. 2008). The earliest surveys had shown that HI is a tracer of spiral structure in galaxies, and the THINGS provides some of the highest quality maps revealing this correlation as well as other characteristics. From a morphological point of view, HI maps tend to reveal (1) enhanced surface brightness in star-forming features such as spiral arms, rings, and pseudorings; (2) extended gaseous disks, such that the HI extent can exceed the optical extent by several times; and (3) supernova-driven and star-formation-driven, windblown holes.

Figure 1-39 shows the HI morphologies of eight THINGS galaxies ranging from the Sab galaxy NGC 4736 to the Sm galaxy DDO 50 (Holmberg II). Bright Sc galaxies like NGC 628 (M74) and NGC 5236 (M83) show HI distributions that extend well beyond the optical disks. These extended patterns can include large spirals as in NGC 628. In M81 and M83, the HI traces the optical spiral structure well. Large rings or pseudorings are seen in NGC 2841 and NGC 4258, while M81 shows an intermediate-scale ring of gas that is much less evident optically. The bright star-forming inner ring in NGC 4736 is well defined and easily distinguished in HI, but the galaxy's diffuse stellar outer ring is more of an asymmetric spiral zone.

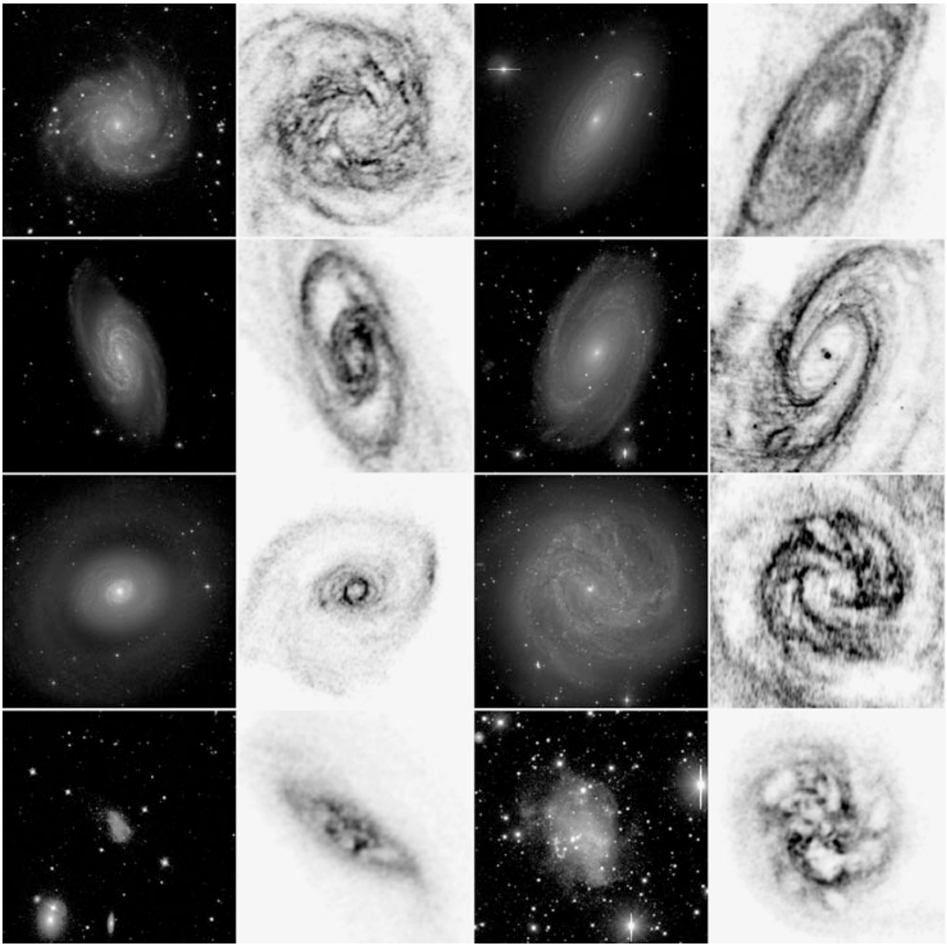


Fig. 1-39

HI morphologies (Walter et al. 2008) of eight galaxies as compared to optical *B*-band morphologies. *Left panels*: NGC 628 (M74), NGC 4258 (M106), NGC 4736 (M94), and DDO 154. *Right panels*: NGC 2841, NGC 3031 (M81); NGC 5236 (M83), and DDO 50 (Ho II)

Most interesting in HI maps are the obvious holes where there appears to be a deficiency of neutral gas compared to surrounding regions. Especially large holes are seen in the HI morphology of the late-type dwarf DDO 50. These holes are thought to be regions cleared by the stellar winds and explosions of massive stars contained or once contained within them. The holes are 100 pc–2 kpc in size and have different systematic properties in early- and late-type galaxies in the sense that holes may last longer in late-type dwarfs owing to the lack of serious shear due to strong differential rotation (Bagetakos et al. 2009). Holes may also be found outside the standard isophotal optical angular diameter.

The dwarf galaxy DDO 154 shown in ➔ *Fig. 1-39* has one of the largest ratios of HI to optical diameter, a factor of 6 at least according to Carignan and Purton (1998), who also estimated that 90% of the mass of the galaxy is in the form of dark matter. The HI and optical morphologies bear little resemblance to one another. Another example of strongly uncorrelated HI and optical morphologies is NGC 2915, a very gas-rich galaxy whose optical morphology is a blue compact dwarf while its at least five times larger HI morphology includes a prominent outer spiral with no optical counterpart (Meurer et al. 1996), leading to the concept of a purely "dark spiral." Bertin and Amorisco (2010) consider a general interpretation of such large outer HI spirals, especially those seen in galaxies like NGC 628 and NGC 6946: the spirals represent short trailing waves that carry angular momentum outward from corotation and, at least in the gaseous component, penetrate a normal barrier at the OLR of the stellar disk pattern and go far out into the HI disk. The short trailing waves are thought to excite the global spiral arms seen in the main optical body of the galaxy (where the prominent optical spirals of NGC 628 and NGC 6946 are found). How NGC 2915 fits into this picture is unclear since the main stellar body of this galaxy is not a grand-design spiral.

Galaxies whose HI disks do not extend much beyond the optical light distribution are also of interest. NGC 4736 in ➲ *Fig. 1-39* is an example. The large, nearby ringed barred spiral NGC 1433 was found by Ryder et al. (1996) to have neutral hydrogen gas concentrated in its central area, its inner ring, and in its outer pseudoring, with no gas outside the visible disk light and a lower amount of gas in the bar region compared to the ring regions. Higdon et al. (1998) showed that in the ringed, barred spiral NGC 5850, neutral gas is concentrated in the inner ring and in the asymmetric outer arm pattern, with little or no emission detected outside this pattern. The asymmetry led Higdon et al. to propose that NGC 5850 has possibly experienced a high-speed collision with nearby NGC 5846.

In galaxy clusters, it is well known that environmental effects can truncate an HI disk so that it is *smaller* than the optical disk light. This is dramatically illustrated in the high-resolution VLA HI maps of Virgo Cluster galaxies by Chung et al. (2009), who found that galaxies within 0.5 Mpc of the cluster core have severely truncated HI disks typically less than half the size of the optical standard isophotal diameter, D_{25}. A variety of earlier studies had already shown these galaxies to be HI deficient compared to field galaxies of similar types. As noted in ➲ Sect. 10.2, an interaction between a cluster galaxy's ISM and the intracluster medium can account for these unusual modifications of HI morphology. Chung et al. also provide evidence for this interaction in some morphologies that appear to be gas stripping in progress.

The Berkeley-Illinois-Maryland Survey of Nearby Galaxies (BIMA SONG; Regan et al. 2001; Helfer et al. 2003) provided some of the highest quality CO maps of normal galaxies. CO emission is often seen in intermediate (Sab-Sd) galaxies, which have a high enough gas content and metallicity to allow the ^{12}CO J = 1 – 0 2.6-mm emission line to be detectable. The CO distributions of eight such galaxies from this survey are shown in ➲ *Fig. 1-40*. These display some of the range of molecular gas morphologies seen. CO traces the inner spiral arms of NGC 628, 1068, and 4535 and is seen along the bar of NGC 2903. A common CO morphology is a large-diameter ring of giant molecular clouds (GMCs), without a central CO concentration, as seen in NGC 2841 and 7331. The rings are the peaks of exponentially declining distributions. The coherent inner molecular gas ring in NGC 7331 appears more like a typical resonance ring and has an estimated molecular gas mass of $3.4 \times 10^9\ M_\odot$ (Regan et al. 2004). The CO distribution in NGC 2403 appears to be concentrated in individual GMCs, with little diffuse emission, while that in NGC 3351 is characterized by a small central bar aligned nearly perpendicular to the galaxy's primary bar. Helfer et al. (2003) show that the Milky Way, M31, and M33 have CO morphologies that are consistent with the range of morphologies found by BIMA SONG.

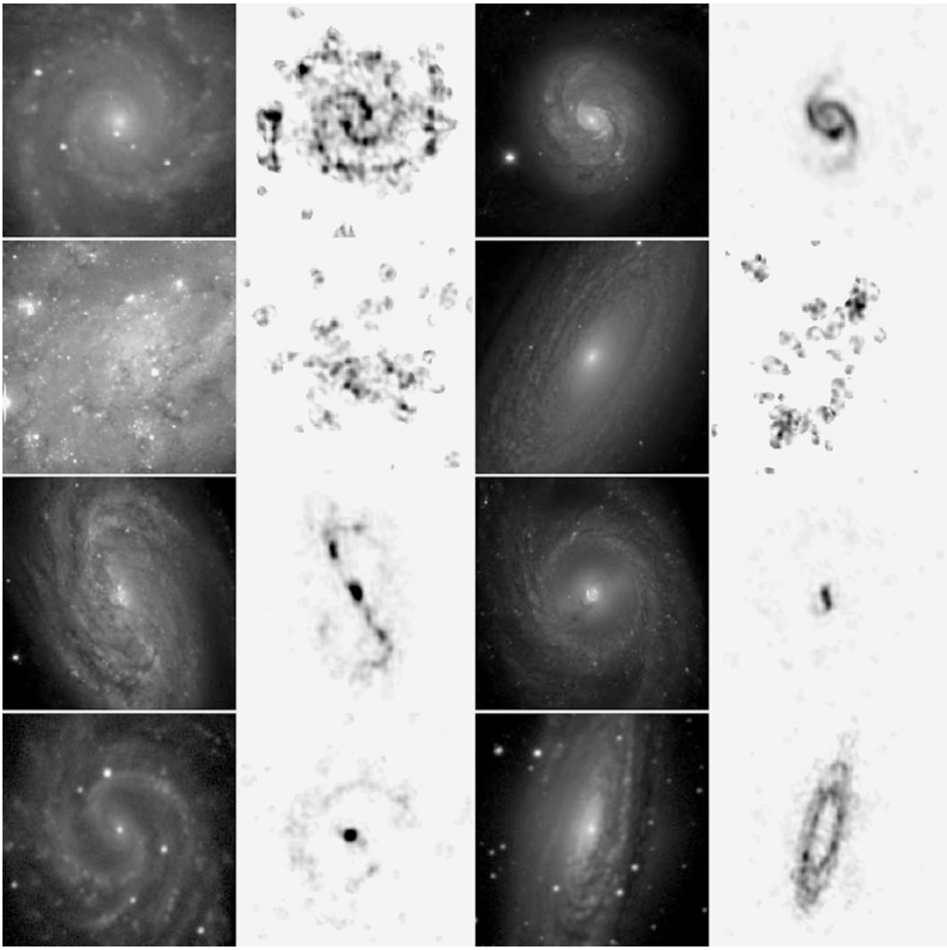


Fig. 1-40
CO morphologies (Helfer et al. 2003) of eight galaxies as compared to optical *B*-band morphologies. *Left panels*: NGC 628 (M74), NGC 2403, NGC 2903, and NGC 4535. *Right panels*: NGC 1068 (M77), NGC 2841; NGC 3351 (M95), and NGC 7331

12 Infrared Observations: Galactic Stellar Mass Morphology

Infrared observations have considerable advantages over optical observations of galaxies. While traditional *B*-band images are sensitive to dust and extinction (both internal and external), the effects of extinction in the near- and mid-IR are much less and become virtually negligible at 3.6 μm. Spiral galaxies imaged at wavelengths successively longer than *B*-band become progressively smoother looking, not only due to the reduced effect of extinction, but also to the de-emphasis of the young blue stellar component. The combination of these effects has led to the popular idea that IR imaging reveals the "stellar backbone" of galaxies, that is, the distribution of actual stellar mass (e.g., Rix and Rieke 1993; Block et al. 1994). Thus, infrared imaging has become a staple for studies of the gravitational potential and stellar mass distribution in galaxies (e.g., Quillen et al. 1994) and for the quantification of bar strength from maximum relative gravitational torques (e.g., Buta and Block 2001; Laurikainen and Salo 2002).

Infrared imaging has also revealed interesting outer structures such as the large outer red arcs seen in M33, which have been interpreted by Block et al. (2004b) to be swaths of extremely luminous carbon stars formed from external accretion of low-metallicity gas. Power spectrum analysis of the IR structures in classical spirals has been used to detect azimuthal "star streams" and to evaluate the role of turbulence on star formation and spiral structure (Block et al. 2009 and references therein).

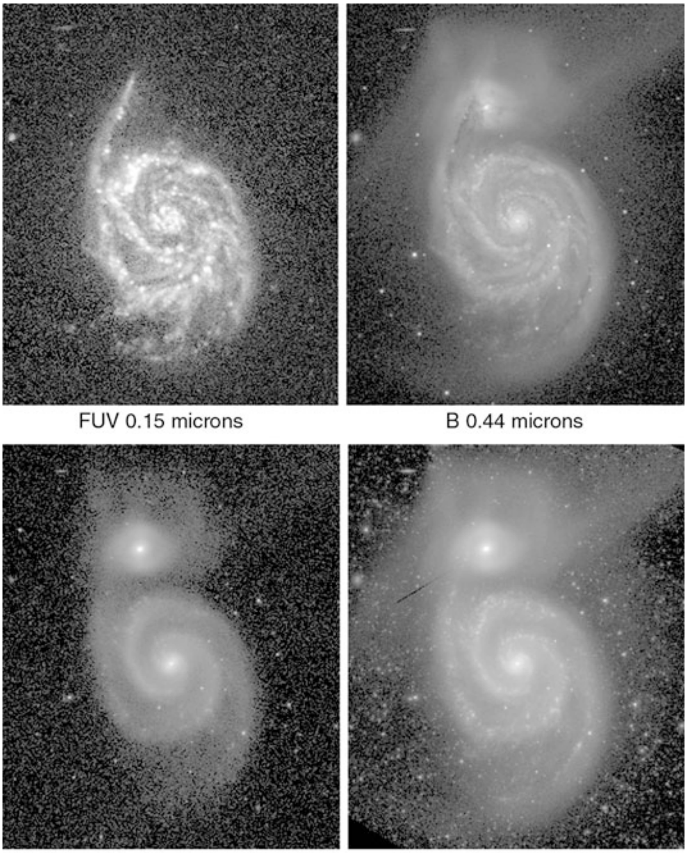
Near-infrared imaging from 0.8 to 2.2 μm can be successfully obtained from ground-based observatories but with the serious drawback that the brightness of the sky background increases substantially over this range. As a result, it has not been possible to achieve a depth of exposure at, for example, 2.2 μm comparable to the kinds of depths achievable at optical wavelengths without excessive amounts of observing time. The first major near-IR survey designed for large-scale morphological studies was the Ohio State University Bright Spiral Galaxy Survey (OSUBSGS, Eskridge et al. 2002), which included optical $BVRI$ and near-IR JHK images of 205 bright galaxies of types S0/a to Sm in a statistically well-defined sample selected to have total blue magnitude $B_T \leq 12.0$ and isophotal diameter $D_{25} \leq 6'.5$. This survey allowed a direct demonstration of how galaxy morphology actually changes from optical to near-IR wavelengths, not merely for a small, selected sample of galaxies, but for a large sample covering all spiral subtypes. The main near-IR filter used in this survey was the H-band at 1.65 μm.

The OSUBSGS was later complemented by the *Near-Infrared S0 Survey* (NIRS0S, Laurikainen et al. 2005, 2006, 2007, 2009, 2010; Buta et al. 2006), a K_s-band imaging survey of 174 early-type galaxies in the type range S0$^-$ to Sa, but mostly including S0s, some of which were misclassified as ellipticals in RC3. NIRS0S images are deeper than OSUBSGS near-IR images owing to the use of larger telescopes and longer on-source times. Although S0 galaxies are dominated by old stars and are usually smooth even in blue-light images, the K_s-band was chosen to complement the OSUBSGS sample of spirals in order to make a fair comparison between bar strengths and bulge properties of S0s and spirals. Also, S0 galaxies are not necessarily dust-free, and near-IR imaging is still necessary to penetrate what dust they have. NIRS0S has led to several important findings about S0 galaxies: (1) a class of S0s, not previously recognized, having prominent lenses but very small bulges that are more typical of Sc galaxies than of earlier type spirals (e.g., NGC 1411, Laurikainen et al. 2006); (2) considerable evidence that S0 galaxies have pseudobulges just as in many spirals (Laurikainen et al. 2007). While the bulges of the latter are likely to be made of rearranged disk material in many cases (➋ Sect. 9), those in S0s are likely to be related to the evolution of bars. S0 bulges tend to be nearly exponential (Sersic index $n \leq 2$), are supported against gravity by rotation rather than random motions, and often include clear inner disks; (3) good correlations between bulge effective radii, r_e, and disk radial scale length, h_R, as well as between the K_s-band absolute magnitudes of the bulge and disk, suggest that S0 bulges are not formed from hierarchical mergers, implying that S0s could be stripped spirals, although the lower bar fraction in S0s suggest that this is in conjunction with evolution due to bars and ovals; (4) 70% of S0-S0/a galaxies have ovals or lenses, suggesting that bars have been weakened in such galaxies over time; and (5) not only bulges but also disks of S0s are similar to those in S0/a-Scd spirals.

The *Two-Micron All-Sky Survey* (2MASS, Skrutskie et al. 2006) provided near-IR JHK_s images of a much larger galaxy sample than either the OSUBSGS or NIRS0S, although these images lack the depth of the OSUBSGS and NIRS0S images in general. 2MASS provided considerable information on near-infrared galaxy morphology, which led to the extensive *2MASS Large Galaxy Atlas* (Jarrett et al. 2003).

The best imaging of galaxies at mid-IR wavelengths has been obtained with the *Spitzer Space Telescope* using the Infrared Array Camera (IRAC, Fazio et al. 2004) and 3.6, 4.5, 5.8, and 8.0 μm filters. The 3.6 and 4.5 μm filters provide the most extinction-free views of the stellar mass distribution in galaxies, while the 5.8 and 8.0 μm filters reveal the interstellar medium (ISM) (Pahre et al. 2004). The loss of coolant in 2008 prevented further observations with the 5.8 and 8.0 μm filters, but the 3.6 and 4.5 μm filters could still be used. This led to the *Spitzer Survey of Stellar Structure in Galaxies* (S^4G, Sheth et al. 2010), a 3.6 and 4.5 μm survey of 2,331 galaxies of all types closer than 40 Mpc. These wavelengths sample the Rayleigh–Jeans decline of the stellar spectral energy distribution of all stars hotter than 2,000 K. S^4G images shown here are from Buta et al. (2010a) and are based on presurvey archival images processed in the same manner as survey images. *Spitzer* observations have a very low background compared to ground-based near-IR observations, and thus IRAC images are the deepest galaxy images ever obtained in the IR.

Figure 1-41 compares images of M51 at four wavelengths: the GALEX 0.15 μm band, the *B*-band (0.44 μm), the near-infrared K_s-band (2.2 μm), and the IRAC 3.6 μm band.



FUV 0.15 microns

B 0.44 microns

K_s 2.2 microns

IRAC 3.6 microns

Fig. 1-41
Multiwavelength images of M51

Only the B- and K_s-band images are from ground-based observations. The GALEX image reveals the extensive star formation in the spiral arms and the complete absence of star formation in the companion NGC 5195 as well as in the complex tidal material north of the companion. The star formation in the arms is more subdued in the B-band and almost completely subdued in the K_s-band. The arms are so smooth in the K_s-band that the galaxy resembles an Sa or Sab system. (The B-band type is Sbc.) Surprisingly, this is not the case in the 3.6 μm image whose considerably greater depth compared to the K_s-band image is evident. The spiral arms in the 3.6 μm band are lined by numerous resolved objects, many of which are correlated with the star-forming regions seen in the B-band. This is dramatically seen also in the SB(s)cd galaxy NGC 1559 (➋ *Fig. 1-42*), where an IRAC 3.6 μm image is compared with a B-band image. These show that resolved features in the deep 3.6 μm image are well correlated with B-band star-forming complexes. Thus, mid-IR 3.6 μm images are *not* completely free of the effects of the extreme population I stellar component (see discussions in Block et al. 2009; Buta et al. 2010a).

A sampling of S^4G images as compared to B-band images for the same galaxies from the dVA is shown in ➋ *Fig. 1-43*. The four galaxies shown, NGC 584, 1097, 628, and 428, have dVA types of $S0^-$, SBb, Sc, and Sm, respectively, thus covering almost the entire Hubble–de Vaucouleurs sequence. Although the very dusty interacting system NGC 1097 looks slightly "earlier" at 3.6-μm, these images show again that on the whole the morphology in the two wavebands is very similar. The same is seen for other galaxies described by Buta et al. (2010a), who found that 3.6 μm types, judged using the same precepts described in the dVA for blue-light images, are well correlated with blue-light types. On average, mid-IR classifications for RC3 S0/a-Sc galaxies are about one stage interval earlier than B-band classifications, with little difference for types outside this range. The correlation is much better than what was expected from

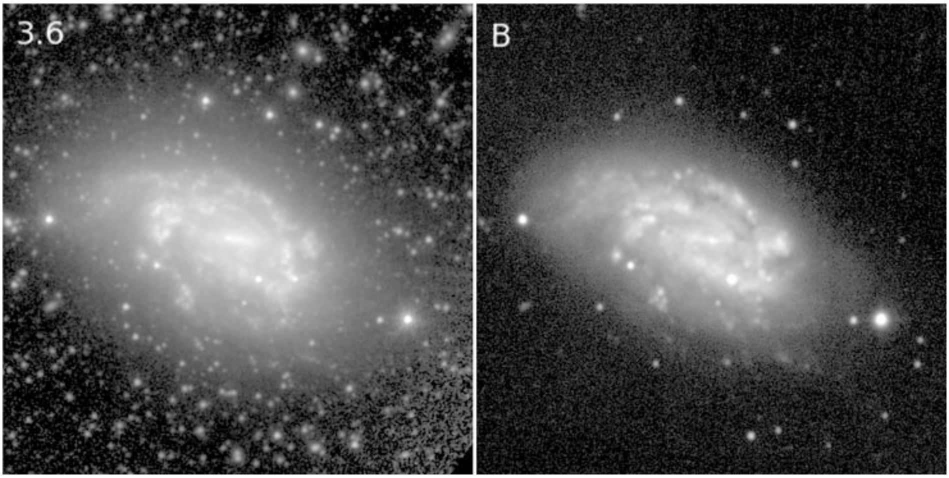


Fig. 1-42

Comparison of IRAC 3.6 μm image of NGC 1559 (*left*) with a ground-based *B*-band image of the same galaxy at *right*. Note the significant correspondence of features between the two very different wavelength domains in this case

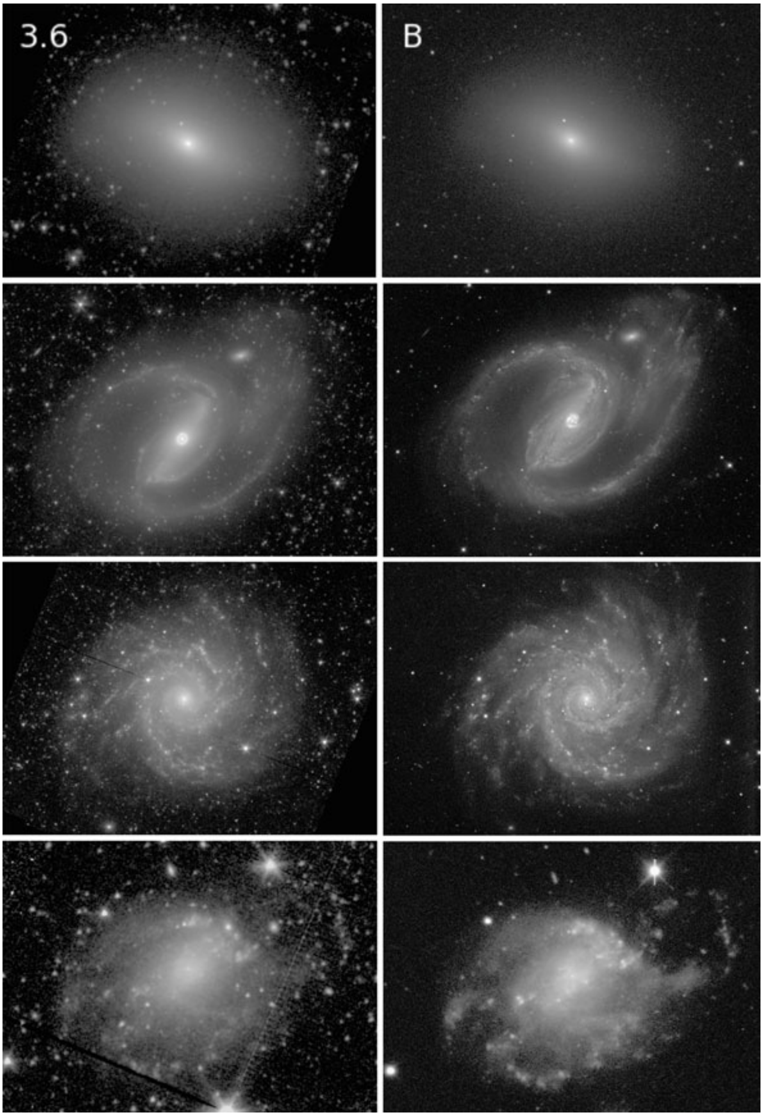


Fig. 1-43

Comparison of IRAC 3.6 μm images (*left frames*) with ground-based *B*-band images for (*top* to *bottom*): NGC 584, 1097, 628, and 428

previous near-IR studies (Eskridge et al. 2002). 3.6 μm galaxy morphology is sufficiently contaminated by recent star formation to allow the same criteria defined for blue-light images to be used for galaxy classification, a surprising result.

Drastic differences between 3.6 μm and *B*-band morphology are seen only for the most dusty galaxies. One example is NGC 5195, shown also in *Fig. 1-41*. This galaxy, classified as Irr II in the Hubble Atlas and as I0 by de Vaucouleurs, appears as a regular early-type galaxy of type SAB(r)0/a (see also Block et al. 1994). Other galaxies that can look very different are

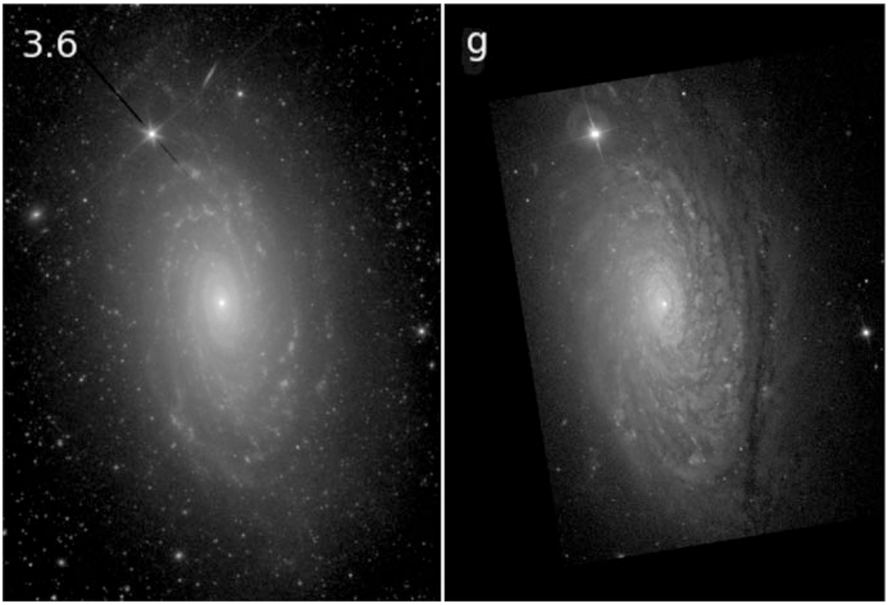


Fig. 1-44
Comparison of IRAC 3.6 μm and SDSS *g*-band images of the flocculent spiral galaxy NGC 5055

flocculent spirals such as NGC 5055 (*Fig. 1-44*). The flocculence largely disappears at 3.6 μm, and a more global pattern is seen (see also Thornley 1996). The type of NGC 5055, Sbc, remains largely unchanged from *B* to 3.6 μm.

As noted by Helou et al. (2004), the mid-IR wavelength domain marks the transition from emission dominated by starlight to emission dominated by interstellar dust. While images at 3.6 μm show the stellar bulge and disk almost completely free of dust extinction, an image at 8 μm shows very little starlight but considerable emission from the ISM in the form of glowing dust.

Figure 1-45 shows an 8 μm image of the nearby spiral galaxy M81 as compared to a $B - I$ color index map coded such that blue star-forming features are dark while red dust lanes are light. The 8 μm image of M81 shows that its ISM is closely associated with its spiral arms. Comparison with the $B - I$ color index shows that both the star-forming arms as well as near-side dust lanes can be seen at 8 μm. Even the far-side lanes in the bulge region are clear at 8 μm, where no tilt asymmetry is manifested. Willner et al. (2004) argue that the dust emission from M81's ISM is likely dominated by polycyclic aromatic hydrocarbons (PAHs; Gillett et al. 1973) which have a prominent emission feature at 7.7 μm. Willner et al. also showed good correspondence between the nonstellar dust emission in M81 and the distribution of near-ultraviolet (NUV) emission. Some regions with bright dust emission and little NUV emission were attributed to excessive UV extinction, while areas with bright NUV and little dust emission were attributed to the effects of supernovae.

The $B - I$ color index map of M81 shows an additional set of dust lanes that have no counterpart in the 8 μm map. These lanes are oriented roughly perpendicularly to the major axis about halfway between the center and the northern arm. Sandage and Bedke (1994) interpret these as foreground dust associated with high galactic latitude nebulosities in the halo of our Galaxy.

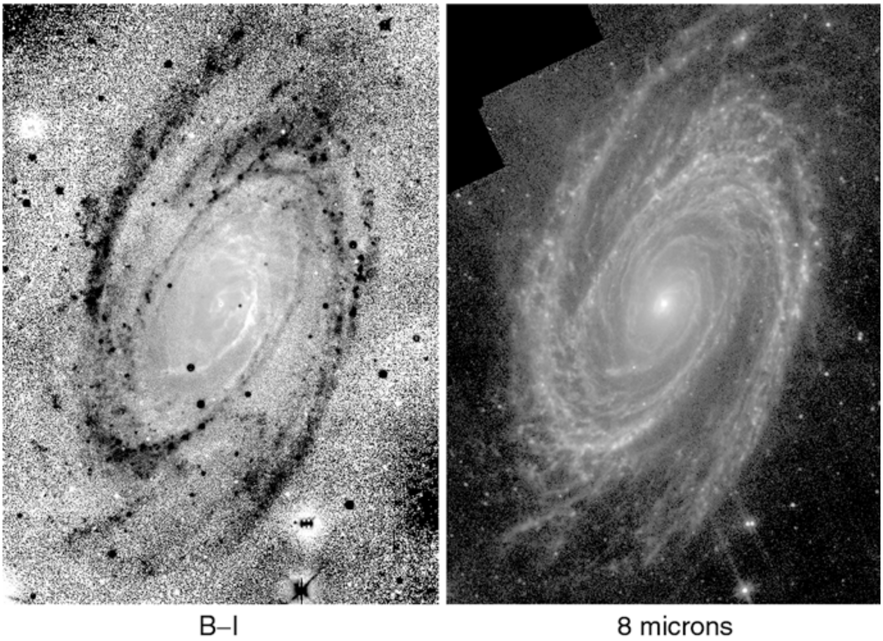


Fig. 1-45
A comparison between a *B* – *I* color index map and an 8 μm dust emission map of the nearby spiral galaxy M81

13 Intermediate and High Redshift Galaxy Morphology

The key to detecting observable evidence for galaxy evolution, that is, to actually see morphological differences that are likely attributable to evolution, is to observe galaxies at high redshift with sufficient resolution to reveal significant details of morphology. Butcher and Oemler (1978) had already found strong evidence for morphological evolution in the excess number of blue galaxies in very distant ($z = 0.4$–0.5) rich galaxy clusters. These authors suggested that the blue galaxies are spiral galaxies and that by the present epoch, these are the galaxies that became the S0s that dominate nearby rich, relaxed clusters (like the Coma Cluster, Abell 1656). An excellent summary of the issues connected with high redshift morphological studies is provided by van den Bergh (1998).

Progress on galaxy morphology at high redshift could only be achieved with the resolution and depth of the Hubble Space Telescope. The Hubble Deep Field North (HDF-N, Williams et al. 1996), South (HDFS, Volonteri et al. 2000), and Ultra-Deep Field (HUDF, Beckwith et al. 2006), and the GOODS (Great Observatories Origins Deep Survey; Giavalisco et al. 2004), GEMS (Galaxy Evolution from Morphology and SEDS; Rix et al. 2004), COSMOS (Cosmic Evolution Survey; Scoville et al. 2007), and other surveys (e.g., Cowie et al. 1995), have provided a large body of information to work with. For example, studies of galaxies in the redshift range $0.3 \le z \le 0.9$ show that the proportion of irregular-shaped galaxies dramatically increases (e.g., Abraham et al. 1996). This means that the Hubble sequence as we know it did not always exist but was built up over time via mergers or secular evolution or both. Observations of submillimeter sources (Chapman et al. 2003) suggest some of these irregulars are extended major mergers.

Interpretation of evolution in the various deep surveys depends on knowledge of redshifts, which can be difficult to measure spectroscopically. A very useful technique for isolating galaxies in high redshift ranges is the UV-dropout method (Steidel and Hamilton 1992). Galaxies are compared in different filters, such as B_{435}, V_{606}, i_{775}, and z_{850}. If the redshift is high enough to move the Lyman limit at 0.0912 μm out of any of the first three filters, there will be a significant drop in flux owing to absorption by hydrogen, and the galaxy is said to "drop out." Galaxies found in this way are called "Lyman break" galaxies because the effect is partly caused by the spectral characteristics of hot stars, which show such a break. (Another cause is UV self-absorption.)

Steidel et al. (1996) used the UV-dropout approach to identify high z galaxies in the HDF-N and also used direct spectroscopy to confirm that the method works. Beckwith et al. (2006) utilized the method to identify galaxies in the HUDF having z from 3.5 to 7. If a galaxy is seen to drop out of a U-band filter such as F300W and seen in a B-band filter such as F450W (both used for the HDF-N; Williams et al. 1996), then the redshift range selected is $z = 2.4$–3.4. Van den Bergh (1998) argues that most Lyman break galaxies are young ellipticals or bulges.

van den Bergh et al. (2000) describe the issues connected with morphological classifications of galaxies to redshifts of $z \approx 1$. Resolution, band shifting, and selection effects due to the magnitude-limited nature of surveys all enter into the interpretation of intermediate to high redshift galaxy morphology. Resolution is important because, typically, a nearby galaxy, will have 100 times or more pixels in the image than a high z galaxy will have for classification. This is fewer pixels than sky survey images of nearby galaxies would have. The bandpass effect is important because the B-band, the standard wavelength for historical galaxy classification, is not sampling the same part of the spectrum as it would for nearby galaxies. For example, at $z = 1$, a B-band filter samples mid-ultraviolet light (≈ 0.22 μm) and would be much more sensitive to young star-forming regions than it would be for nearby galaxies. Ideally, then, for comparison with nearby galaxies we would like to choose a redshifted band as close as possible to the *rest-frame* B-band. Even accounting for all these effects, significant differences between nearby and distant morphologies do exist. For example, van den Bergh et al. (2000) discuss the paucity of grand-design spirals and barred galaxies in the HDF-N and use artifically redshifted images of nearby galaxies to demonstrate that the deficiencies are likely to be real.

Figure 1-46 shows several of the different categories of intermediate and high redshift galaxy morphologies, based on V and i-band images from the GEMS and Hubble UDF. The redshifts range from 0.42 to 3.35 and provide a wide range of look-back times. First, in such a range, some galaxies look relatively normal, as shown by the spiral and elliptical galaxies in the two upper left frames of *Fig. 1-46*. The $z = 0.59$ spiral is classifiable as type SA(s)bc and the elliptical as type E3. The $z = 0.99$ spiral shown in the middle right frame of *Fig. 1-46* has larger clumps, no clear central object, and more asymmetry than the $z = 0.59$ spiral but is still recognizable as a spiral. However, other less familiar categories are found. In general, high redshift galaxies have smaller linear diameters than nearby galaxies on average (e.g., Elmegreen et al. 2007a [EEFM07]).

"Chain galaxies" were first identified by Cowie et al. (1995) and are linear structures with superposed bright knots that have sizes and blue colors similar to normal late-type galaxies and relatively flat major axis luminosity profiles. They have the shapes of edge-on disk galaxies but lack clear bulges or nuclei. A recent study by Elmegreen et al. (2004a) [EES04] of faint

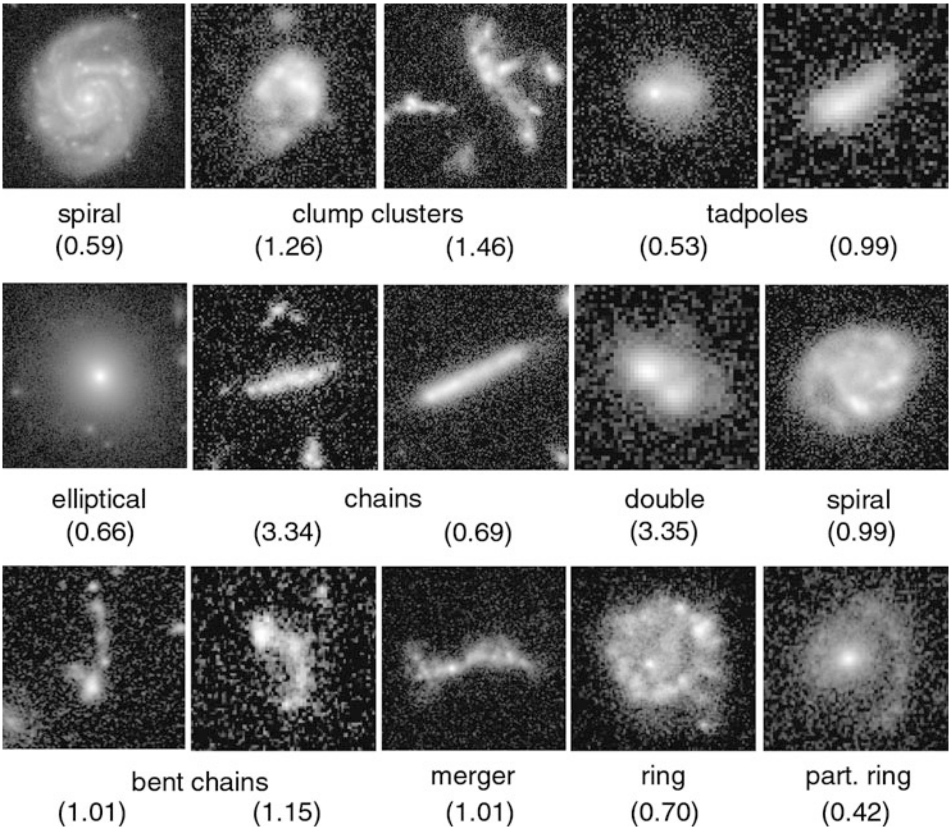


Fig. 1-46

Intermediate to high redshift galaxy morphology (*V* and *i*-bands). The categories are due to EES04 and EE06. The number in parentheses below each frame is the redshift *z* of the galaxy shown

galaxy morphologies (redshifts 0.5–2) in the Advanced Camera for Surveys (ACS) "Tadpole" galaxy (UGC10214) field showed that chain galaxies are the most common linear morphology at magnitudes fainter than $I = 22$, accounting for more than 40% of the sample. Their dominance (also found by Cowie et al. 1995) is interpreted by EES04 as a selection effect because relatively optically thin edge-on galaxies are more favored to be seen near the limit owing to a higher projected surface brightness than for face-on versions of the same galaxies. EES04 suggest that chain galaxies are edge-on irregular galaxies that will evolve to late-type disk galaxies. Chains are the most flattened linear morphology at faint magnitudes.

"Clump clusters" (EES04) are somewhat irregular collections of blue knots or clumps with very faint emission between clumps. The clumps have sizes of ≈500 pc and masses of $\approx 10^8$–$10^9\ M_\odot$. Both of the examples shown in *Fig. 1-46* have $z > 1$. Elmegreen et al. (2004b) [EEH04] identified clump clusters as the face-on counterparts of the linear chain galaxies, based on the similarities of the properties of the clumps with those seen in chain galaxies and on the distribution of axis ratios of the systems as compared with normal disk galaxies. The lack of a bulge clump is also consistent with this conclusion. Nevertheless, analysis of NICMOS IR images in the HUDF led Elmegreen et al. (2009a) to conclude that 30% of clump clusters and 50% of chain galaxies show evidence of young bulges, implying that at least half of these galaxies might be genuinely bulgeless. In a related study, Elmegreen et al. (2009b) show that the best local analogues of clump clusters are dwarf irregular galaxies like Ho II, scaled up by a factor of 10–100 in mass. This study also brought attention to clump clusters with faint red background disks, as opposed to blue clump clusters which lack such a feature. Elmegreen et al. argue that the red background clump clusters are part of an evolutionary sequence leading from the blue clump clusters to spirals with a "classical" bulge (e.g., KK04). The clumps, formed by gravitational instabilities in a turbulent disk, are large and few in number and thus will eventually coalesce near the galaxy center if they survive the effects of supernova explosions (e.g., Elmegreen et al. 2008).

"Tadpoles" (van den Bergh et al. 1996) are asymmetrically shaped "head-tail" morphologies with a bright off-centered nucleus and a tail, like a tadpole. A rare local example is NGC 3991. Tadpoles were recognized in 3% of the galaxies in HDF-N and were found to be very blue in color. Usually both the head and the tail are blue, but van den Bergh et al. (1996) show one example where the head is red and the tail is blue. EEH04 showed that tadpoles have neither exponential major axis profiles nor clear bulges, and in their sample of linear objects, tadpoles are the least frequently seen.

The bottom frames of ➋ *Fig. 1-46* show bent chains and rings or partial rings (Elmegreen and Elmegreen 2006 [EE06]). The rings and partial rings are thought to be mostly collisional in nature (i.e., like the conventional ring galaxies shown in ➋ *Fig. 1-27*) and show the different morphologies expected when small companions plunge through a larger disk galaxy in different ways (Appleton and Struck-Marcell 1996). Although bent chains resemble the partial rings, they lack offset nuclei and any evidence of a background, more face-on disk. EE06 suggest that bent chains are simply warped versions of the more common linear chains that have suffered an interaction. The ages of the bent chain clumps are younger than those found in rings and partial rings, and EE06 argue that relative separations and sizes of the clumps indicate they form by gravitational instabilities.

Elmegreen et al. (2005) show that approximately one third of the ellipticals catalogued in the HUDF have prominent blue clumps in their centers (see also Menanteau et al. 2001, 2004). They argued that these clumps probably imply accretion events based on comparison of their magnitudes and colors with local field objects. Menanteau et al. (2001) were able to reproduce the color distributions with a model having a starburst superposed on a preexisting older stellar population.

Galaxy morphology at intermediate and high redshifts also includes obvious interacting cases as well as possible merger morphologies. Bridges, tidal tails, plumes, and even M51 analogues are seen as in nearby galaxies, but are smaller in scale than for nearby objects (EEFM07). The middle frame of the bottom row shows a possible merger in progress of two bent chains (or, alternatively, two interacting spirals), called an "assembly galaxy" by EEFM07 because they appear to be assembling from smaller objects. EEH04 and EEFM07 also discuss the double systems, considered another category of the linear systems. The double systems like the $z = 3.35$ one shown in ➋ *Fig. 1-46* are probably merging ellipticals. EEFM07 also describe "shrimp galaxies," which appear to be interacting galaxies with a single curved arm or tail, curling at one end into a "body."

Other studies of high redshift galaxy morphology have focused on the specific redshift ranges that are selected by the UV-dropout technique. Conselice and Arnold (2009) examined the morphologies of galaxies in the $z = 4$–6 range from the HUDF and measured quantitative parameters such as the concentration-asymmetry-clumpiness (star formation) parameters (CAS; Conselice 2003) and other related parameters that are useful for distinguishing mergers from normal galaxies. The CAS system is based on simple global parameters that are easily derived automatically for large numbers of galaxies. Conselice (2003) tied the C parameter to the past evolutionary history of galaxies while parameters A and S measure more active evolution from mergers and star formation. Conselice and Arnold found that half of the HUDF dropout galaxies they studied have significant asymmetries and may be undergoing merging, while the other half is mainly smooth symmetric systems that may have collapsed quickly into a temporary, quiescent state.

Other quantitative approaches to these issues include the Sersic n index that characterizes radial luminosity profiles (Ravindranath et al. 2006; Elmegreen et al. 2007b) and the Gini coefficient (Abraham et al. 2003; Lotz et al. 2006). The Gini coefficient provides a way of quantifying high redshift morphology that does not depend on galaxy shape or the existence of a well-defined center and is well suited to the kinds of objects shown in ❯ *Fig. 1-46*. Lotz et al. (2006) found in a sample of 82 Lyman break galaxies that 10–25% are likely mergers, 30% are relatively undisturbed spheroids, and the remainder are disks, minor mergers, or postmergers.

Given the rise in peculiar and irregular-shaped galaxies with increasing redshift, the question naturally arises: when did the Hubble sequence and all its accompanying details fall into place? This question is considered by Conselice et al. (2004), who quantitatively analyzed a well-defined high redshift sample using the CAS system. Conselice et al. identify "luminous diffuse objects" (LDOs) as galaxies having C less than 1σ below the average, and "luminous asymmetric objects" (LAOs) as galaxies having $A > S$. Some of both classes of objects are covered by the Elmegreen et al. categories described above. All of the LDOs and LAOs have $M_B < -19$, and Conselice et al. suggest such objects might be the precursors of modern disk and elliptical galaxies. These are found in the redshift range $0.2 < z < 2$, suggesting the present-day Hubble sequence began taking shape in this interval. Conselice et al. (2008) consider the morphologies of galaxies more massive than $10^{10}\ M_\odot$ and in the range $1.2 < z < 3$. To a z_{850} magnitude of 27, the majority of these galaxies are peculiar. They conclude that such galaxies undergo 4.3 ± 0.8 mergers to $z = 3$.

14 Giant Low-Surface-Brightness Galaxies

The van den Bergh luminosity classes highlight how luminosity and surface brightness generally go together. Low surface brightness usually means low luminosity and small size, hence a dwarf classification. However, the discovery of rare giant low-surface-brightness (GLSB) galaxies by Bothun et al. (1987) shows that morphology can sometimes be misleading for judging absolute luminosity. The hallmarks of these objects are a relatively normal bulge and an extremely low-surface-brightness, very large disk. Disk radial scale lengths and luminosities are unusually large, and extrapolated disk central surface brightnesses are unusually faint compared to more normal spirals. The disks tend to be relatively smooth with a few large, isolated HII regions. Bothun et al. (1987) point to a model whereby the disks of these galaxies have such a low gas surface density that they are largely unevolved due to the inefficiency of star formation.

GLSB galaxies can be classified within the Hubble–Sandage and de Vaucouleurs classification systems although, as noted by McGaugh et al. (1995), the majority are classified later than stage Sc. Bulges, bars, rings, and spiral patterns are evident in some examples in spite of the low

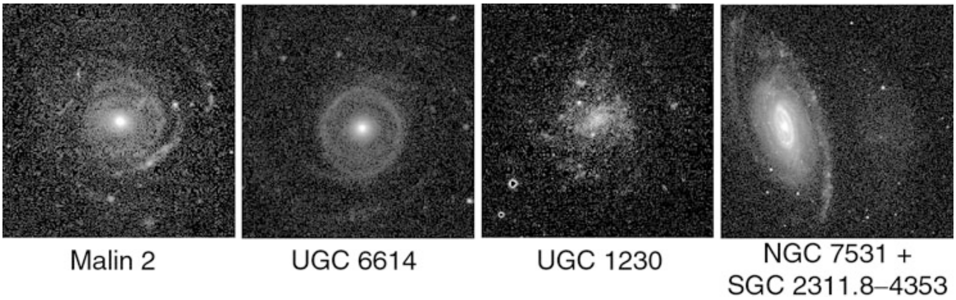


Fig. 1-47
Examples of giant or large low-surface-brightness galaxies. In the far-*right panel*, SGC 2311.8–4353 is the diffuse object to the *right* of high-surface-brightness spiral NGC 7531. All of these images are *B*-band

disk surface brightness. *Figure 1-47* shows three of the originally recognized GLSB examples (McGaugh et al. 1995): Malin 2 (also known as F568–6), UGC 6614, and UGC 1230. In the images, the length of a side is 131, 38, and 77 kpc, respectively. These can be compared to the giant normal spiral NGC 7531 in the far-right frame, where the length of a side is also 38 kpc. Malin 2 and UGC 6614 are especially enormous physical objects. UGC 1230 is also very large for such a late-type morphology. van den Bergh (1998) likens the size of Malin 2 to the core of a cluster of galaxies. He considers "monsters" like Malin 1 and Malin 2 to be only one of three types of LSB galaxies. Some LSB galaxies are as big as normal galaxies, like UGC 1230. Most LSB galaxies, however, are dwarfs: ellipticals, irregulars, and, less frequently, spirals. These are described further in Sect. 15.2.

An example of another object that could be considered a large LSB galaxy, but which lacks a bulge or any evident recent star formation, is SGC 2311.8–4353, the mysterious ghostlike companion close to the right of NGC 7531 in *Fig. 1-47*. This peculiar object is two-thirds the size of NGC 7531 (Buta 1987) but has unknown redshift. If it is associated with NGC 7531, it would be as much as 30 kpc in diameter at the faintest detectable isophote level and would clearly not be a dwarf. A recent extremely deep image of this pair by Martínez-Delgado et al. (2010) suggested to these authors that SGC 2311.8–4353 is a cloud of tidal debris from a disrupted satellite that resembles the "umbrella" structures seen in hierarchical cosmological simulations.

A recent study of three GLSBs (Malin 1, UGC 6614, and UGC 9024) by Rahman et al. (2007) showed that IR emission from such objects is consistent with their optically determined low star formation rates, with the diffuse optical disks being undetected from two of the three. A dynamical study of two GLSBs (Malin 1 and NGC 7589) by Lelli et al. (2010) led the authors to conclude that at least in these cases, the GLSB galaxy can be thought of as an inner high-surface-brightness galaxy having a very extended LSB disk. This is based on the steeply rising rotation curves found for these galaxies, which is very much like what is seen in early-type high-surface-brightness galaxies.

Impey and Bothun (1997) argue that LSB galaxies brighter than $M_B = -14$ contribute significantly to the luminosity density of the local universe, are dark matter dominated at almost all radii, and have an evolutionary history involving late collapse of a low amplitude perturbation, a low star formation rate, and very slow changes. Large LSBs are greatly underrepresented in galaxy catalogues but are clearly an important class of objects.

15 Galaxy Morphology in Color

15.1 Normal Galaxies

The Sloan Digital Sky Survey (Gunn et al. 1998; York et al. 2000) provides the largest body of information on the colors of galaxies. The multiwavelength imaging in *ugriz* filters has allowed the production of high-quality color images for thousands of galaxies. Although the Hubble–Sandage–de Vaucouleurs classification systems were based on blue-light images alone, it is still possible to reliably classify galaxies with SDSS color images, which are based on combined *gri* images (Lupton et al. 2004). In such images, one can directly see the stellar population differences that characterize different galaxy types.

Figure 1-48 shows the color Hubble sequence from E to Sm. Once galaxy colors were systematically measured using photoelectric photometry (de Vaucouleurs et al. 1976), it was noted that integrated colors vary smoothly with advancing stage along the Hubble sequence. The latest stages have corrected total color index $(B - V)^{\circ}_{T} \approx 0.3$–0.4 while E and S0 galaxies have $(B - V)^{\circ}_{T} \approx 0.9$–1.0 (e.g., Buta et al. 1994). The latter colors correspond to yellow-orange while the former are bluish white. The colors begin to change at S0/a and Sa and become progressively bluer. *Figure 1-48* shows the reason for the change. Galaxies earlier than Sa are dominated by old stars having the colors of K giants. As stage advances from Sa to Sm, the spiral structure becomes progressively more important compared to the bulge. Since the arms are dominated by complexes of massive young stars, this makes the integrated colors of the galaxies become progressively bluer until by the end of the sequence, the bluer colors of these stars have overcome the yellowish light of the background old disk stars. *Figure 1-48* also shows that the intermediate colors of intermediate types such as Sb and Sc are due to the yellowish orange colors of bulges and bars as combined with the bluer colors of spiral arms.

The analysis of integrated SDSS galaxy colors for more than 100,000 galaxies led to one of the most dramatic findings of the survey: a clear bimodality in the distribution of color that correlates with morphology: a red peak that includes mainly E, S0, and Sa galaxies and a blue peak that includes mainly Sb, Sc, and Irr galaxies (Strateva et al. 2001). Although the correlation of galaxy color with types had been known for a long time from photoelectric measurements (e.g., de Vaucouleurs 1961; Buta et al. 1994), the large sample provided by SDSS allowed the bimodality to be demonstrated to a high degree of significance. In plots of $u - r$ color index versus absolute M_r magnitude, the galactic equivalent of a stellar H-R diagram, nearby early-type galaxies follow a narrow band called the red sequence, while nearby later-type, mostly star-forming galaxies appear as a broad blue sequence (also sometimes called the "blue cloud"). Baldry et al. (2004) showed that the bimodality (in the form of a double Gaussian number distribution over all types) is detectable from $M_r \approx -15.5$ to $M_r \approx -23$, being undetectable only for the most luminous galaxies. Wyder et al. (2007) showed that use of GALEX near-ultraviolet magnitudes and optical r-band magnitudes provides even greater discrimination between the Gaussian components. Bell et al. (2004) showed that the bimodality is detectable in faint galaxies to $z \approx 1$, indicating that this characteristic of the galaxy population extends to a look-back time of at least 9 Gyr. It is thought that galaxies evolve from the blue sequence to the red sequence as their star formation is quenched, perhaps through mergers, gas depletion, or AGN feedback (e.g., Martin et al. 2007). The possibility of evolution has engendered great interest in the galaxies lying near the minimum of the bimodal distribution (the so-called green valley; Thilker et al. 2010 and references therein).

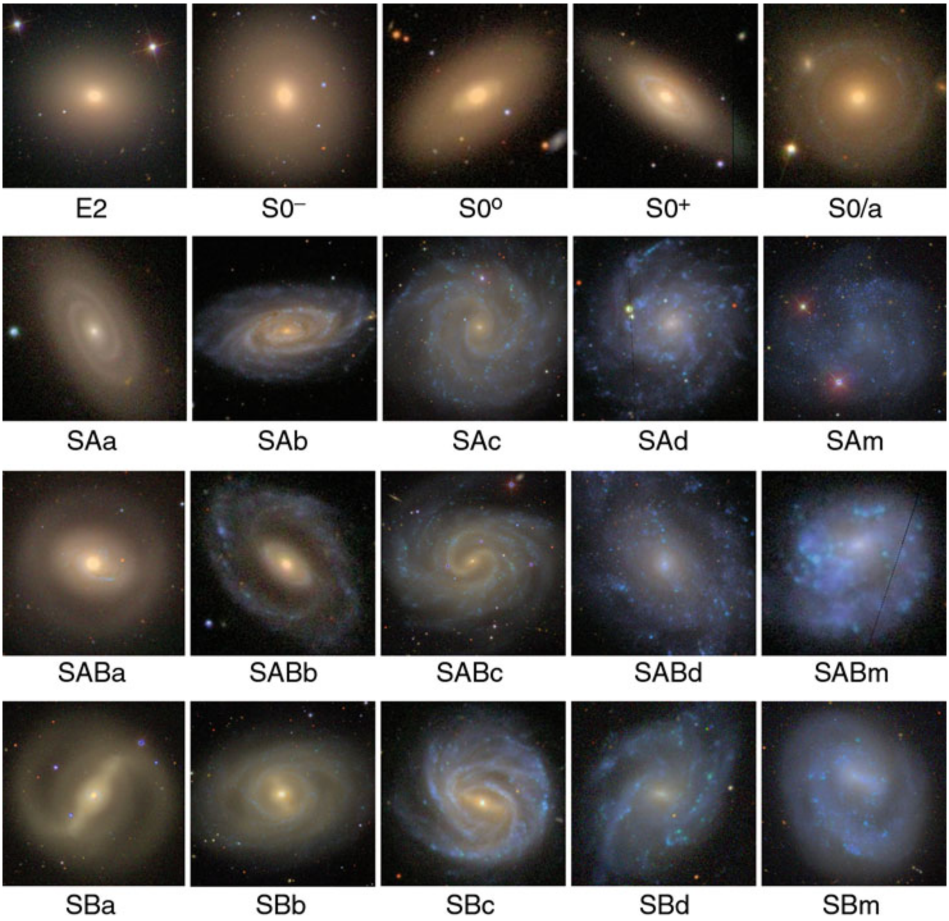


Fig. 1-48
The Hubble tuning fork of ellipticals, S0s, and spirals of different bar classifications are shown here using SDSS color images. The galaxies are (*left* to *right*): *Row 1* – NGC 3608, 4203, 6278, 4324, and 932. *Row 2* – NGC 4305, 5351, 3184, 5668, and IC 4182. *Row 3* – NGC 4457, 5409, 4535, 5585, and 3445. *Row 4* – NGC 4314, 3351, 3367, 4519, and 4618

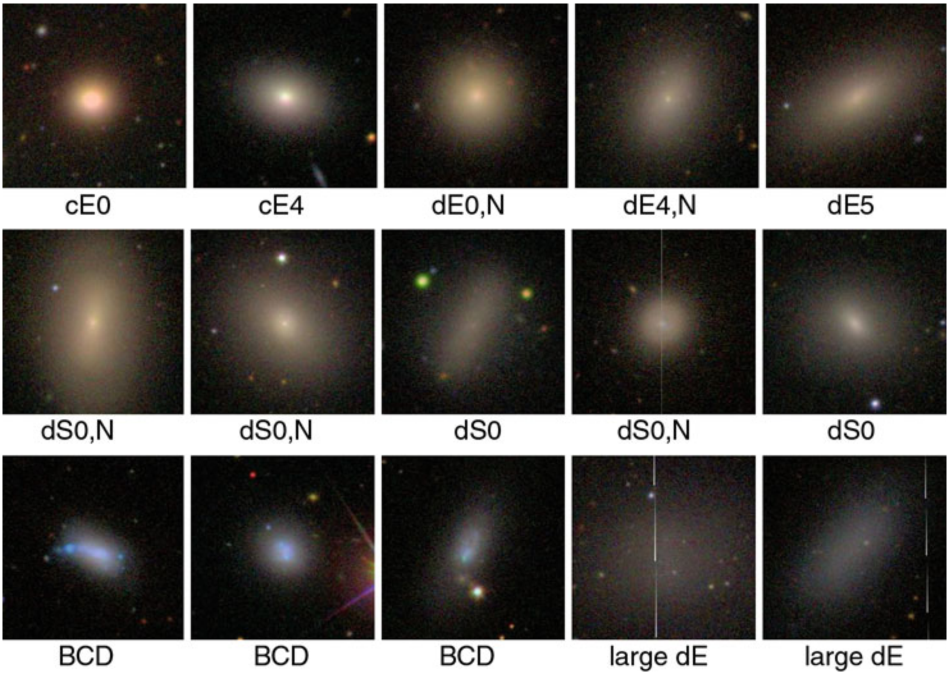
15.2 Dwarf Galaxies

Virtually all the galaxies shown in *Fig. 1-48* are of relatively high luminosity, with absolute blue magnitudes M_B° averaging about −20. When physical parameters such as M_B° are considered, it becomes clear that the peculiar shape of the de Vaucouleurs classification volume shown in *Fig. 1-3* only highlights the morphological diversity of families and varieties at each stage but does not tell us about the *physical parameter space* at each stage, which expands considerably at each end of the volume (McGaugh et al. 1995). Most known dwarf galaxies are either early- or late-type, but not intermediate.

15.2.1 dE, dS0, BCD, and cE Galaxies

The most extensive study of dwarf galaxy morphologies was made by Binggeli et al. (1985) [BST], who used deep photographs to probe the low-luminosity population of the Virgo Cluster using mostly morphology to deduce cluster membership. Examples of several categories of Virgo Cluster dwarf galaxies are shown in ➲ *Fig. 1-49* using SDSS color images and the classifications of BST. The most common type is the dwarf elliptical, or dE type, which accounts for 80% of the galaxies in the BST catalogue. dE galaxies range from $M_B = -18$ to -8 (Ferguson and Binggeli 1994). Many dEs have an unresolved, starlike nucleus whose presence is indicated by an N attached to the type, as in dE0,N. The top right panels of ➲ *Fig. 1-49* show three fairly typical examples. Possibly related to these normal dE systems are the larger, lower surface brightness ellipticals ("large dE") shown in the two lower right frames of ➲ *Fig. 1-49*.

The second row of ➲ *Fig. 1-49* shows examples of the interesting class of dwarf S0 galaxies. All of the examples shown are distinct from dEs in showing a smooth structure but with additional features such as a lens or a weak bar. dS0 galaxies can also be nucleated and are called dS0,N. In addition to the low-surface-brightness dEs and dS0s, the Virgo Cluster includes two high-surface-brightness classes of dwarfs. The cE category refers to compact ellipticals that resemble M32.



◘ Fig. 1-49
Examples of dwarf galaxies in the Virgo Cluster, drawn from the catalogue of Binggeli et al. (1985) and highlighted using SDSS color images. The classifications are from the BST catalogue and the galaxies are (*left* to *right*): *Row 1* – NGC 4486B, IC 767, IC 3470, IC 3735, and UGC 7436. *Row 2* – NGC 4431, IC 781, IC 3292, VCC 278, IC 3586. *Row 3* – VCC 459, VCC 2033, VCC 841, IC 3475, and IC 3647

The blue compact dwarf (BCD) galaxies are a special class of star-forming dwarf irregulars characterized by a few bright knots imbedded in a stellar background of low surface brightness (Sandage and Binggeli 1984). The most extreme cases are nearly stellar (Thuan and Martin 1981). The knots are often super star clusters associated with 30-Doradus-like HII regions, and the faint background can be very blue (Thuan et al. 1997). Spectroscopically, BCDs have narrow emission lines superposed on a blue continuum, and the lines indicate a low metallicity. ➋ *Figure 1-49* shows three examples from the BST Virgo Cluster catalogue. Gil de Paz et al. (2003) present an atlas of more than 100 BCDs that highlight their structure.

As described in ➋ Sect. 5.1, the dE galaxies shown in ➋ *Fig. 1-49* are *not* the low-luminosity extension of more luminous ellipticals. These together with the dS0 class are labeled "spheroidal" galaxies by Kormendy et al. (2009; see ➋ *Fig. 1-6*), who confirmed the finding by Kormendy (1985) that these galaxies are more related to evolved low-luminosity spirals and irregulars than to genuine ellipticals. Based on correlations of well-defined photometric parameters (e.g., central surface brightness or velocity dispersion vs. core radius or absolute magnitude), these authors link ellipticals like M32 to the actual low-luminosity end of the E galaxy sequence (see also Wirth and Gallagher 1984). Thus, the cE galaxies shown in ➋ *Fig. 1-49* are in a sense truer "dwarf ellipticals" than the dE galaxies shown. In spite of this revised viewpoint, the BST classifications still have value since these are purely morphological interpretations.

15.2.2 Local Group Dwarf Spheroidals and Irregulars

The lowest luminosity galaxies that we can study in detail are in the Local Group. Fortunately, a few are in the area covered by the SDSS so that they can be illustrated in color. These objects, all fainter than $M_V = -12$, are shown in ➋ *Fig. 1-50*. Leo I and Leo II are usually called dwarf spheroidal, or dSph, galaxies (e.g., Ferguson and Binggeli 1994). dSph galaxies tend to have to have absolute visual magnitudes $M_V > -15$ and a low degree of flattening; they are believed to be the most abundant type of galaxy in the universe. Leo I ($M_V = -11.9$) and Leo II ($M_V = -9.6$) look different in part because of their different star formation histories. dSph and dwarf irregular (dI) galaxies are now known to have complex and varied star formation histories that may involve multiple episodes of star formation and effects of interactions (Mateo 1998 and references therein). The other two galaxies in ➋ *Fig. 1-50*, Leo A and DDO 155, are dwarf irregulars having $M_V \approx -11.5$. A detailed HST study of the stellar content of Leo A (Cole et al. 2007)

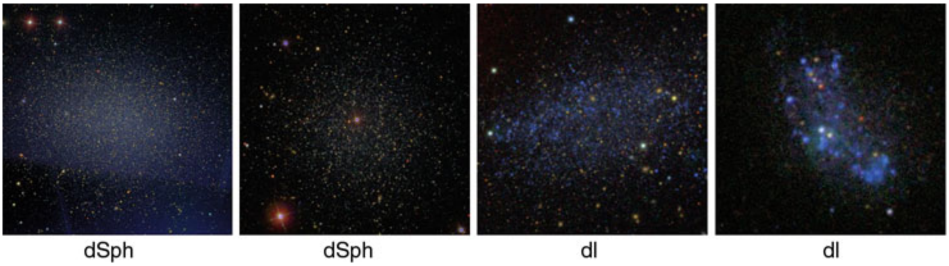


Fig. 1-50
Four Local Group dwarfs having $M_V > -12$ (*left* to *right*): Leo I, Leo II, Leo A, and DDO 155 (All SDSS color images)

showed that 90% of the star formation in the galaxy occurred less than 8 Gyr ago, with a peak at 1.5–3 Gyr ago. A useful summary of the properties of dSph galaxies is provided by van den Bergh (1998) and, most recently, by Tolstoy et al. (2009), who also discuss the star formation histories of Local Group dI galaxies.

The SDSS has facilitated the discovery of many new Local Group dwarf galaxies (Belokurov et al. 2007). For example, the SDSS led to the discovery of one of the faintest known dSph galaxies, a new dwarf in Ursa Major (called the UMa dSph) having $M_V \approx -7$ (Willman et al. 2005). Most interesting is Leo T, which is an $M_V = -7.1$ dSph galaxy with some recent star formation, providing one of the most dramatic illustrations of the link between dSph and dI galaxies, and the least luminous galaxy known to have recent star formation (Irwin et al. 2007).

15.2.3 Dwarf Spirals

Sandage and Binggeli (1984) described the classification of dwarf galaxies based on the Virgo Cluster and concluded that there are "no real dwarf spirals." This refers mainly to dwarf spirals that might be classified as types Sa, Sb, or Sc, that is, having both a bulge and a disk. Dwarf late-type spirals are already built into de Vaucouleurs's modified Hubble sequence as Sd-Sm types and connect directly to Magellanic irregulars, as shown in Fig. 1 of Sandage and Binggeli (1984). Thus, any genuine examples of dwarf Sa, Sb, or Sc spirals would be of great interest as they would challenge the idea that for a galaxy to be able to make well-defined spiral arms, it would have to be more massive than some lower limit (estimated as $5 \times 10^9\ M_\odot$ by Sandage and Binggeli).

Four of the best cases of genuine dwarf spirals are illustrated in ➲ *Fig. 1-51*. The two rightmost frames show IC 783 (BST type dS0,N) and IC 3328 (BST type dE1,N), both Virgo Cluster members having absolute magnitudes $M_B^\circ \approx -16$ to -17 and found to have subtle spiral structure by Barazza et al. (2002) and Jerjen et al. (2000), respectively. The patterns are hard to see in the direct SDSS color images shown in ➲ *Fig. 1-51*, but these authors use photometric models, Fourier decomposition, and unsharp-masking to verify the reality of the patterns. Barraza et al. conclude that many of the bright early-type dwarfs in the Virgo Cluster have disks. An example with a bar *and* spiral arms is NGC 4431 (shown as dS0 in ➲ *Fig. 1-49*).

The leftmost panel of ➲ *Fig. 1-51* shows an HST wide V-band image (Carollo et al. 1997) of NGC 3928, an absolute magnitude $M_B^\circ = -18$ galaxy which on small-scale, overexposed images looks like an E0 but which harbors a miniature (2-kpc diameter), low-luminosity spiral

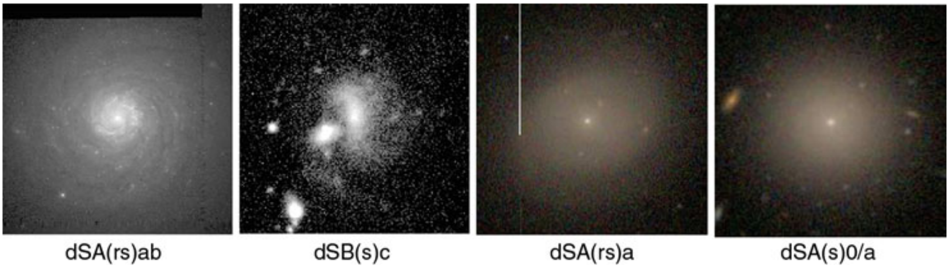


Fig. 1-51
Four dwarf spiral galaxies (*left* to *right*): NGC 3928, D563–4, IC 783, and IC 3328

(van den Bergh 1980b). Based on spectroscopic analysis, Taniguchi and Watanabe (1987) have suggested that NGC 3928 is a spheroidal galaxy which experienced an accretion event that supplied the gas for star formation in the miniature disk.

Schombert et al. (1995) brought attention to possible dwarf field spirals. One of their examples, D563–4, is shown in the second frame from the left in ➋ *Fig. 1-51*. This galaxy has $M_B^\circ \approx -17$. A few other examples are given in the paper, and Schombert et al. find that they are not in general grand-design spirals, are physically small, and have low HI masses. However, van den Bergh (1998) considers all of Schombert et al.'s examples as subgiant spirals rather than true dwarf spirals. A possible true dwarf spiral given by van den Bergh is DDO 122 (type S V).

15.3 Galaxy Zoo Project

The Galaxy Zoo project (Lintott et al. 2008) has made extensive use of SDSS color images. The project uses a website to enlist the help of citizen scientists worldwide to classify a million galaxies as well as note interesting and unusual cases in various forum threads. With such a large database to work from, and the potential for discovery being real, the project has attracted many competent amateur galaxy morphologists. One such discovery was a new class of galaxies called "green peas," which are starlike objects that appear green in the SDSS composite color images (➋ *Fig. 1-52*, left). Cardemone et al. (2009) used auxiliary SDSS data to show that peas are galaxies that are green because of a high equivalent width of [OIII] 5007 emission. They are sufficiently distinct from normal galaxies and quasars in a two-color $g - r$ versus $r - i$ plot that such a plot can be used to identify more examples. Other characteristics noted are that peas are rare, no bigger than 5 kpc in radius, lie in lower density environments than normal galaxies but may still have morphological characteristics driven by mergers, are relatively low in mass and metallicity, and have a high star formation rate. Cardemone et al. conclude that peas are a distinct class of galaxies that share some properties with luminous blue compact galaxies and UV-luminous high redshift galaxies.

Another prominent colorful galactic-sized object identified by Galaxy Zoo is "Hanny's Voorwerp" (➋ *Fig. 1-52*), a peculiar collection of blue clumps just south of IC 2497.

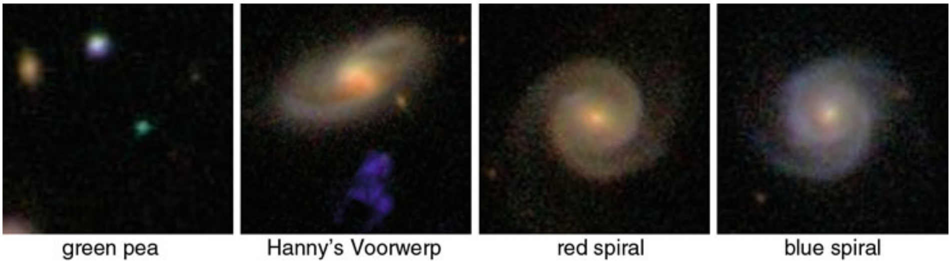


Fig. 1-52
Interesting morphologies and color characteristics found by the Galaxy Zoo project team participants. A "green pea" is a star-like galaxy with a high flux in [OIII] 5007. "Hanny's Voorwerp" is a cloud of ionized gas that may be the light echo of a quasar outburst in the nucleus of nearby IC 2497. Red spirals are morphologically similar to blue spirals but have a lower star formation rate (All SDSS color images)

Lintott et al. (2009) found that the Voorwerp is mostly ionized gas, and after ruling out some possible sources of ionizing radiation, concluded that the object could be the first identified case of a quasar light echo. The implication is that the companion galaxy underwent a temporary quasar phase.

Masters et al. (2010) examine the properties of face-on late-type spiral galaxies whose colors are much redder than is typical (➲ *Fig. 1-52*, right panels), suggesting that they are passive objects where star formation has largely ceased or is lower than normal. These authors showed that red spirals are not excessively dusty and tend to be near the high end of the mass spectrum. A range of environmental densities was found, implying that environment alone is not sufficient to make a spiral red. A significantly higher bar fraction was found for red spirals as compared to blue spirals, which Masters et al. suggest could mean the bars themselves may have acted to shut down the star formation in these galaxies.

A major philosophical aspect of Galaxy Zoo is the value of human visual interpretation of galaxy morphology. That is, the human eye can integrate the detail in an image more reliably than a computer program can, in spite of the latter's ability to classify numbers of galaxies well beyond the capability of a single individual. This philosophy was used to compile a major catalogue of galaxy merger pairs described by Darg et al. (2010). In the initial setup of Galaxy Zoo, a single button allowed a classifier to select whether an object was a "merger" based simply on the appearance of peculiarities. For each galaxy, a weighted average number, f_m, was derived that characterized the fraction of classifiers who interpreted a pair of galaxies as a merger, essentially "morphology by vote" and in a way reminiscent of RC3 where morphological types were in many cases based on a weighted average from a small number of classifiers (Buta et al. 1994). Taking $f_m > 0.4$, Darg et al. (2010) identified 3,003 pairs and groups of merging galaxies and also showed that the spiral to elliptical galaxy ratio in merger pairs is a factor of 2 higher than for the global galaxy population, suggesting that mergers involving spirals are detectable for a longer period than those that do not involve spirals.

A similar philosophy to Galaxy Zoo was used by Buta (1995) to compile the Catalogue of Southern Ringed Galaxies, and also by Schawinski et al. (2007), who visually classified 48,023 SDSS galaxies to identify a significant-sized sample of early-type galaxies for a study of the connection between nuclear activity and star formation. In the latter study, visual interpretation was argued to be needed to avoid bias against star-forming early-type galaxies which would be excluded from color-selected samples. A major result of this study was the identification of a time sequence whereby an early-type galaxy has its star formation suppressed by nuclear activity, a manifestation of AGN feedback.

15.4 Isolated Galaxies

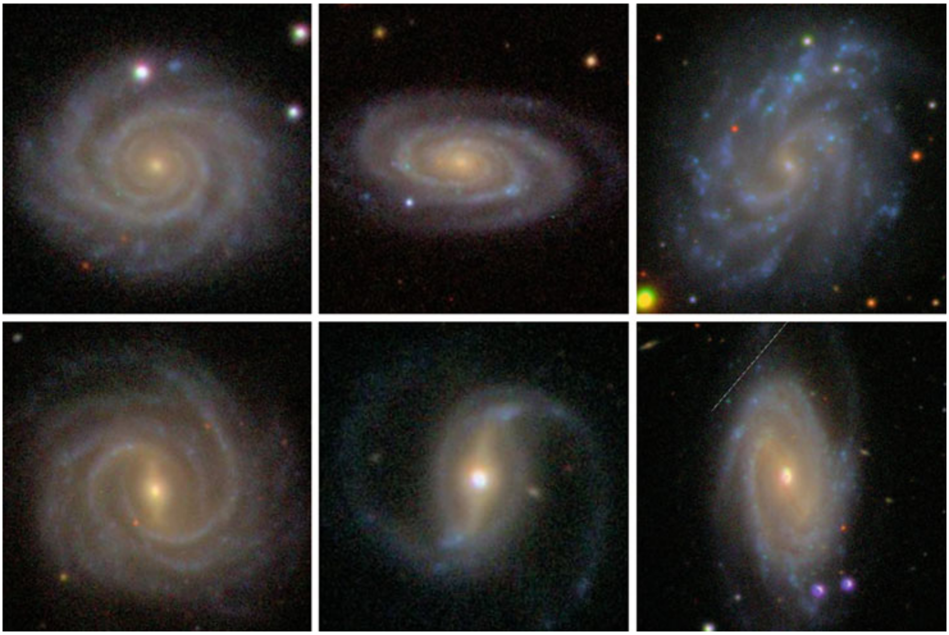
If interactions and mergers can have profound effects on galaxy morphology, then the morphology of *isolated* galaxies clearly is of great interest. Such galaxies allow us to see how internal evolution alone affects morphology, that is, what pure "nature" morphologies look like. Karachentseva (1973) compiled a large catalogue (the Catalogue of Isolated Galaxies, or CIG) of 1,050 isolated galaxies that has proven very useful for examining this issue. A galaxy of diameter D is suggested to be isolated if it has no comparable-sized companions of diameter d between $D/4$ and $4D$ within a distance of $20d$. Verdes-Montenegro et al. (2005) show that this means that an isolated galaxy 25 kpc in diameter, in the presence of a typical field galaxy velocity of 150 km s^{-1}, has not been visited by a comparable mass companion during the past 3 Gyr.

These authors discuss the limitations of the CIG (e.g., the isolation criteria do not always work), but in general it is the best source of isolated galaxies available.

Sulentic et al. (2006) examined all 1,050 CIG galaxies on Palomar II sky survey charts in order to refine the sample and found that isolated galaxies cover all Hubble types. Of these, 14% were found to be E/S0 types, while 63% were Sb-Sc types, with the spiral population more luminous than the E/S0 population. Over the type range Sa-Sd, the proportion rises to 82%. Thus, an isolated galaxy sample is very spiral rich. Nevertheless, the presence of early-type galaxies in the sample implies that these are not likely to be "nurture" formed, as such galaxies might be in denser environments. The refinement of the CIG sample forms the basis of the Analysis of the Interstellar Medium of Isolated Galaxies (AMIGA) project (Verdes-Montenegro et al. 2005).

➲ *Figure 1-53* shows six examples of isolated Sb-Sc galaxies from the AMIGA sample, based on SDSS color images. All of these look relatively normal, but it is interesting how nonbarred galaxies like NGC 2649 and 5622 (upper left frames in ➲ *Fig. 1-53*) show such conspicuous global spirals, which argues that the spirals in these galaxies have not been excited by an interaction.

Durbala et al. (2008) analyzed the photometric properties of 100 isolated Sb-Sc AMIGA galaxies and found that a majority have pseudobulges rather than classical bulges. In comparing the properties of isolated galaxies with a sample of Sb-Sc galaxies selected without an isolation criterion, Durbala et al. found that isolated spirals have longer bars and, using CAS parameters, also less asymmetry, central concentration, and clumpiness.



◘ Fig. 1-53

SDSS color images of six isolated Sb-Sc galaxies, both barred and nonbarred, from the AMIGA sample, a refined version of the Catalogue of Isolated Galaxies (Karachentseva 1973). The galaxies are (*left* to *right*): *Row 1*: NGC 2649, 5622, and 5584; *Row 2*: NGC 4662, 4719, and 2712

Durbala et al. (2009) analyzed the Fourier properties of the same set of 100 isolated spirals and estimated bar lengths and strengths. Earlier types in the sample were found to have longer and higher contrast bars than later types. Spiral arm multiplicities were investigated also, and it was shown that cases having an inner $m = 2$ pattern and an outer $m = 3$ pattern occurred in 28% of the sample. Elmegreen et al. (1992) argued that in such morphologies, the $m = 3$ pattern is driven internally by the $m = 2$ pattern and that three-armed patterns measure the time elapsed since an interaction.

15.5 Deep Field Color Imaging

Particularly interesting in the domain of color galaxy morphology are the various deep field surveys that have been at the heart of high redshift studies. The most recent is the HST Wide Field Camera 3 Early Release Science (ERS) data (Windhorst et al. 2011), which provides very deep panchromatic images based on ten filters ranging from 0.2 to 2 μm in wavelength. ❯ *Figure 1-54* shows three subsections of the main ERS field, which coincides with the GOODS south field (Giavalisco et al. 2004). The galaxies seen range in redshift from $z = 0.08$ to at least $z = 3$. Windhorst et al. (2011) argue that images like the ERS field are deep enough to allow probing of galaxy evolution in the crucial redshift range $z \approx 1$–3 where the galaxies assembled into their massive shapes. By $z \approx 1$, the Hubble sequence was largely in place. ❯ *Figure 1-54* shows a variety of interesting nearby as well as high z morphologies, including some of those illustrated in ❯ *Fig. 1-46*. Images like the ERS field allow us to connect the local and distant galaxy populations in unprecedented detail.

16 Large-Scale Automated Galaxy Classification

The process of galaxy classification thus far described is a manual exercise where an observer attempts to sort a galaxy into its appropriate stage, family, variety, outer ring classification, etc., by visual inspection alone. For a small number of observers, this has been done for as many as 14,000 galaxies by Nair and Abraham (2010) and 48,000 galaxies by Schawinski et al. (2007), while for a large number of observers working in concert (e.g., the Galaxy Zoo project), it has been done for a million galaxies. But critical to such ventures is the preparation of images for classification and a need for a homogeneous, objective approach to very large numbers of galaxies. This has led to extensive application of automated methods for classifying galaxies. For example, Nair and Abraham (2010) visually classified galaxies into coded bins for the purpose of training an automatic classification algorithm.

A variety of approaches have been used, ranging from "artificial intelligence" (neural network) techniques (Odewahn et al. 1992; most recently Banerji et al. 2010) to the CAS approach (Conselice 2003) and the Gini coefficient (Abraham et al. 2003), and to algorithms often used in other disciplines such as the automated cell morphology code described by Shamir (2009). The idea is to first classify a sample of galaxies by eye, use the algorithm to also classify them, and then compare how well the algorithm results agree with the visual results. Often the goal with automatic classification is not to automatically derive classifications such as "(R′)SB(r,nr)ab" but simple distinctions of galaxies such as ellipticals, spirals, S0s, or edge-ons. Conselice et al. (2004) in fact argue that automated classification should not focus on reproducing detailed de

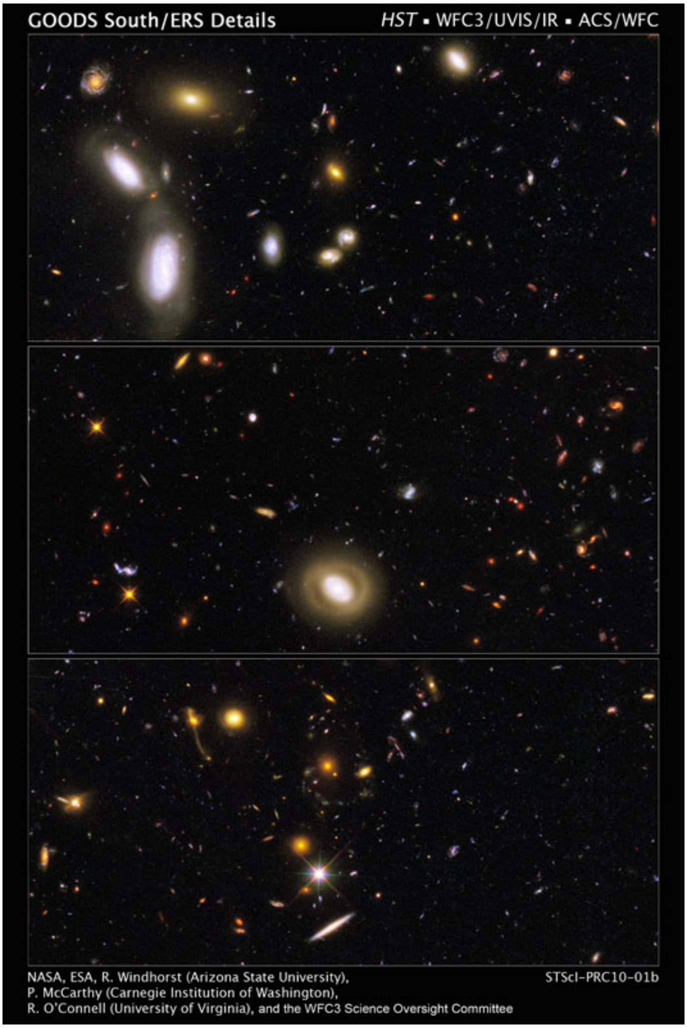

GOODS South/ERS Details
HST • WFC3/UVIS/IR • ACS/WFC
NASA, ESA, R. Windhorst (Arizona State University),
P. McCarthy (Carnegie Institution of Washington),
R. O'Connell (University of Virginia), and the WFC3 Science Oversight Committee
STScI-PRC10-01b

Fig. 1-54
Three subsections from the WFC3-ERS survey of the GOODS-south field (Windhorst et al. 2011). The colors are based on images obtained with ten filters ranging from 0.2 to 2 μm

Vaucouleurs types, for example, but instead focus on more tangible properties like the degree of central concentration, the degree of asymmetry, and the degree of patchiness, all of which in a way can distinguish late-type galaxies from early-type galaxies (but not necessarily ringed or barred galaxies from those lacking such features). The reason for such a view is that high redshift galaxy morphology usually does not have the benefit of the detail seen in nearby galaxies. Van den Bergh (1998) discusses other issues connected with computer classifications, including the limitations of artificial neural networks and the usefulness of objectively measured parameters.

17 The Status and Future of Morphological Studies

The large number of illustrations in this chapter attests to the richness of the diversity of galaxy morphology. It is, of course, not possible to illustrate all aspects of morphology that might be worth discussing, but most interesting is how far *physical galaxy morphology* has come in the 35 years since Allan Sandage wrote his review of galaxy morphology in Volume IX of *Stars and Stellar Systems*. Galaxy morphology is no longer the purely descriptive subject it once used to be.

Internal perturbations such as bars are apparently capable of generating a great deal of the interesting structure we see in disk galaxies, and more theoretical and observational studies should elucidate this further. The impact of bars on morphology seems well understood, as summarized in detail by KK04 in their monumental review article on pseudobulges. Bars redistribute angular momentum and reorganize gas clouds to flow into resonance regions and fuel star formation. The gathering of gas into resonance regions can drive the formation of rings and, indeed, can also build up the central mass concentration to the point of bar destruction. Even failing this, the pileup of gas into the nuclear region can lead to the formation of a pseudobulge. The richness of barred galaxy morphology attests to the strong role secular evolution plays in structuring galaxies.

Progress in understanding the role of mergers and interactions on galaxy morphology has also proceeded at a rapid pace. Great success in numerical simulations and the theory of interacting galaxies has made it possible to link a specific type of interaction to a specific morphology (e.g., collisional ring galaxies). The complex structure of early-type galaxies, with boxy and disky isophote shapes, shells and ripples, and other features, shows that interactions and mergers play an important role in molding galaxies (see the excellent review by Schweizer 1998). With the advent of the Hubble Space Telescope, this role has been elucidated even more clearly because the merger rate was higher in the past.

In spite of the theoretical progress, it is interesting that classical morphology has not lost its relevance or usefulness even after more than 80 years since Hubble published his famous 1926 paper. No matter how much progress in understanding the physical basis for morphology is made, there is still a need for the ordering and insights provided by classical Hubble–Sandage–de Vaucouleurs galaxy classifications. Morphology went through a low phase in the 1980s and 1990s when it was perceived that galaxy classification placed too much emphasis on unimportant details and was too descriptive to be useful. It was thought that the Hubble classification had gone as far as it can go and that another approach needed to be tried to build a more physical picture of galaxies. At that time, there was a sense that the focus should be more on the component "building blocks" of galaxies or what might be called galactic subsystems (e.g., Djorgovski 1992). Quantification of morphology became more possible as advanced instrumentation allowed more detailed physical measurements to be made. In the end, as morphology became better understood, it also became clear what a type such as "(R)SB(r)ab" might really mean, which enhanced the value of classification (KK04). In addition, numerical simulations became sophisticated enough to make predictions about morphology (e.g., the R_1 and R_2 subclasses of outer rings and pseudorings). These types of things, as well as the movement of morphology from the photographic domain to the digital imaging domain, the broadening of the wavelength coverage available to morphological studies from the optical to the ultraviolet and infrared domains, the Sloan Digital Sky Survey, and the accessibility of high redshift galaxies to unprecedentedly detailed morphological study, all played a role in bringing galaxy morphology to the forefront of extragalactic research.

Even so, the writing of this chapter has shown that many important galaxies and classes of galaxies have not been studied well enough to have much modern data available. For example, in spite of the considerable interest in collisional ring galaxies the past 20 years or so, Struck (2010) was forced to lament that ring galaxies "are underobserved." The same can be said for resonance ring galaxies, giant low-surface-brightness galaxies, dwarf spirals, Magellanic barred spirals, counterwinding spirals, and other classes of interacting galaxies. The most that can be said about this is that further studies will likely be made, especially if instrumentation facilitates the objects in question. Rotation and dynamics are far short of photometry for most classes of galaxies but would add a great deal of insight if obtainable.

At the other extreme, early-type (E and S0 galaxies) continue to be the focus of major photometric, kinematic, and theoretical research projects. Important clues to the formation and evolution of such galaxies are contained in their intrinsic shapes (oblate, prolate, triaxial), in the ages, metallicities, and radial mass-to-light ratios of their stellar populations, in their three-dimensional orbital structure, and in the kinematic peculiarities often found in such systems (de Zeeuw et al. 2002). Among the most recent studies are the massive photometric analysis of early-types in the Virgo Cluster by Kormendy et al. (2009) and the ATLAS^{3D} project described by Cappellari et al. (2011). ATLAS^{3D} is the largest kinematic database of high-quality two-dimensional velocity field information ever obtained for early-type galaxies, including 260 such galaxies in a well-defined and complete sample. This survey is simply the latest part of the long-term effort by many researchers, beginning in the 1980s, to understand early-type galaxies in terms of quantitative parameters that can be tied to theoretical models. Early-types have been a persistent enigma in morphological studies, and considerable evidence suggests that the E, S0 sequence as defined by Hubble, Sandage, and de Vaucouleurs hides a great deal of important physics associated with these objects. The ATLAS^{3D} project was designed to exploit the λ_R parameter described by Emsellem et al. (2007; see ➋ Sect. 5.1), which separates early-types into fast and slow rotators and discriminates galaxies along the red color sequence (➋ Sect. 15.1).

Another interesting aspect of early-type galaxies is the frequent presence of nuclear stellar disks, features from tens to hundreds of parsecs in diameter. These have been quantitatively studied recently by Ledo et al. (2010), who argued that such disks could not survive a major merger and their presence thus provides information on the assembly history of early-type systems. Ledo et al. found that such disks are present in 13–23% of early-types, and are rarer in the most massive early-types, are most frequent in the absolute magnitude range $M_B = -18$ to -20 and that there is little evidence for a correlation of the structures with galactic environment.

These advances for early-type galaxies do not mean that quantitative analyses of later-type galaxies are lacking. As codes for two-dimensional photometric decomposition become ever more sophisticated (e.g., Peng et al. 2010; Laurikainen et al. 2010 and references therein), parameters that characterize the bulges, disks, bars, lenses, rings, and spiral patterns are being derived for large numbers of galaxies (especially barred galaxies) that were not reliably decomposable with earlier one-dimensional approaches.

For the future, it is to be hoped that the Sloan Digital Sky Survey will be extended to cover the whole sky and provide access to high-quality morphological studies of several million more galaxies, some of which might have new and exotic structures. The *James Webb Space Telescope* should be able to carry the HST's torch to greater depths and resolutions of high redshift galaxies to further enhance our understanding of galaxy evolution.

Note added in manuscript: Since this chapter was completed, the placement of S0 galaxies in the Hubble tuning fork has seen its most serious challenge. As noted by van den Bergh (1976, 1998), the properties of S0s do not support the idea of their being a transition class between ellipticals and Sa, SBa galaxies. Van den Bergh proposed a "parallel sequence" classification where S0s fall on a sequence S0a, S0b, and S0c parallel to the tuning fork sequence Sa, Sb, Sc. The parallel S0s would represent environmentally transformed spiral types of different bulge-to-total-luminosity ratios. The idea did not take hold partly because of the lack of any known cases of S0c galaxies, that is, very small bulge S0s, and because the relative importance of galaxy transformation was still being evaluated both theoretically and observationally (the "nature vs. nurture" controversy). There was also the general "rule" that galaxy classification should be independent of any particular theoretical idea (e.g., Sandage 2004, 2005).

Now, three studies have led to basically the same conclusion: the correct placement of S0s is in a sequence parallel to spirals, not at the juncture of the tuning fork. Kormendy and Bender (2012) proposed that Sph galaxies are environmentally transformed Magellanic irregular galaxies and belong at the end of the parallel S0 sequence. Thus, parallel sequence classification provides a natural "home" for the enigmatic Sph galaxies that were long thought to be dwarf versions of elliptical galaxies. The kinematic parameter λ_R from the ATLAS^{3D} project also led Cappellari et al. (2011) to conclude that the correct placement of S0s is parallel to spirals and that the best way to order early-type galaxies is kinematically, not morphologically. Finally, both Laurikainen et al. (2011) and Kormendy and Bender (2012) found examples of likely S0c galaxies, thus removing one of the early objections to the parallel sequence idea. Also, Laurikainen et al. (2011 and references therein) used parameter correlations to conclude that S0s are transformed spirals.

None of this completely negates the value of tuning fork classifications, because types such as En, S0$_1$, S0$_2$, S0$_3$ or E$^+$n, S0$^-$, S0°, and S0$^+$ are still morphologically valid visual categories. A galaxy may be classified as type SA(rl)0$^+$ and S0b, and the two classifications tell us something different and complementary. The revisions to the tuning fork represent the coming to fruition of quantitative and interpretive galaxy morphology.

This article is dedicated to Allan Sandage (1926–2010), one of the twentieth century's greatest astronomers, who helped set the stage for galaxy morphology to be one of the most active fields of modern extragalactic research. It was Dr. Sandage's efforts that firmly cemented Hubble's ideas on morphology into astronomy. The author is grateful to Dr. Sandage for the inspiration he provided for this chapter and for his standard of excellence in astronomy.

The author also gratefully acknowledges the helpful comments and suggestions from the following people that considerably improved this article: Martin Bureau, Adriana Durbala, Debra Elmegreen, William Keel, Jeffrey Kenney, Johan H. Knapen, Rebecca Koopmann, John Kormendy, Eija Laurikainen, Barry Madore, Karen L. Masters, Patrick M. Treuthardt, Sidney van den Bergh, and Xiaolei Zhang. The author also thanks Debra Elmegreen for providing the images of high redshift galaxies shown in ➲ *Fig. 1-46*, Masfumi Yagi for the illustrations in ➲ *Fig. 1-29*, and John Kormendy, Jason Surace, Donald P. Schneider, and Rogier Windhorst for the use of published illustrations from specific papers. This chapter uses images from many sources too numerous to acknowledge here but mainly drawn from the dVA, the NASA/IPAC Extragalactic Database, the Ohio State University Bright Spiral Galaxy Survey (OSUBSGS), the Sloan Digital Sky Survey, and several published papers by other authors. The NASA/IPAC Extragalactic Database (NED) is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Funding for the OSUBSGS was provided by grants from the NSF (grants AST 92-17716 and AST 96-17006), with additional funding from the Ohio State University. Funding for the creation and distribution of the SDSS Archive has been provided by the Alfred P. Sloan

Foundation, the participating institutions,, NASA, NSF, the U.S. Department of Energy, the Japanese Monbukagakusho, and Max Planck Society. Observations with the NASA/ESA *Hubble Space Telescope* were obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under contract NAS 5-26555. The *Spitzer Space Telescope* is operated by the Jet Propulsion Laboratory, California Institute of Technology, under NASA contract 1407. The *Two Micron All-Sky Survey* is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. GALEX is a NASA mission operated by the Jet Propulsion Laboratory. GALEX data is from the Multimission Archive at the Space Telescope Science Institute (MAST). Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX09AF08G and by other grants and contracts. This chapter has also made use of THINGS, "The HI Nearby Galaxy Survey" (Walter et al. 2008), and BIMA-SONG, the Berkeley-Illinois-Maryland Survey of Nearby Galaxies (Helfer et al. 2003).

References

Abraham, R. G., van den Bergh, S., Glazebrook, K., Ellis, R. S., Santiago, B. X., Surma, P., & Griffiths, R. E. 1996, ApJS, 107, 1

Abraham, R. G., van den Bergh, S., & Nair, P. 2003, ApJ, 588, 218

Aguerri, J. A. L., Méndez-Abreu, J., & Corsini, E. M. 2009, A&A, 495, 491

Appleton, P., & Struck-Marcell, C. 1996, Fundam Cosm Phys, 16, 111

Arp, H. 1966, ApJS, 14, 1

Arp, H. C., & Madore, B. F. 1987, Catalogue of Southern Peculiar Galaxies and Associations (Cambridge: Cambridge University Press)

Arribas, S., Bushouse, H., Lucas, R. A., Colina, L., & Borne, K. D. 2004, AJ, 127, 2522

Athanassoula, E. 1992, MNRAS, 259, 328

Athanassoula, E. 2003, MNRAS, 341, 1179

Athanassoula, E. 2005, MNRAS, 358, 1477

Athanassoula, E., & Bosma, A. 1985, ARA&A, 23, 147

Athanassoula, E. et al. 1990, MNRAS, 245, 130

Athanassoula, E., Romero-Gómez, M., & Masdemont, J. J. 2009a, MNRAS, 394, 67

Athanassoula, E., Romero-Gómez, M., Bosma, A., & Masdemont, J. J. 2009b, MNRAS, 400, 1706

Bacon, R. et al. 2001, MNRAS, 326, 23

Bagetakos, I., Brinks, E., Walter, F., de Blok, W. J. G., Rich, J, W., Usero, A., & Kennicutt, R. C. 2009, in The Evolving ISM in the Milky Way and Other Galaxies, eds. K. Sheth, A. Noriega-Crespo, J. Ingalls, & R. Paladini, http://ssc.spitzer.caltech.edu/mtgs/ismevol,E17

Bagley, M., Minchev, I., & Quillen, A. C. 2009, MNRAS, 395, 537

Bahcall, J. N., Kirhakos, S., Saxe, D. H., & Schneider, D. P. 1997, ApJ, 479, 642

Baldry, I. K., Glazebrook, K., Brinkmann, J., Ivezić, Z., Lupton, R. H., Nichol, R. C., & Szalay, A. S. 2004, ApJ, 600, 681

Banerji, M. et al. 2010, MNRAS, 406, 342

Barazza, F. D., Binggeli, B., & Jerjen, H. 2002, AJ, 124, 1954

Barazza, F. D., Jogee, S., & Marinova, I. 2008, ApJ, 675, 1194

Barway, S., Khembavi, A., Wadadekar, Y., Ravikumar, C. D., & Mayya, Y. D. 2007, ApJ, 661, L37

Barway, S., Wadadekar, Y., & Khembavi, A. K. 2011, MNRAS, 410, L18

Beckwith, S. V. W. et al. 2006, AJ, 132, 1729

Bell, E. et al. 2004, ApJ, 608, 752

Belokurov, V. et al. 2007, ApJ, 654, 897

Bershady, M., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645

Bertin, G., & Amorisco, N. C. 2010, A&A, 512, 17

Bertin, G., Lin, C. C., Lowe, S. A., & Thurstans, R. P. 1989, ApJ, 338, 78

Bertola, F. 1987, in IAU Symp. 127 (Dordrecht: D. Reidel Publishing Co.), p. 135

Binggeli, B., Sandage, A., & Tammann, G. A. 1985, AJ, 90, 1681

Binney, J. 1992, ARA& A, 30, 51

Blakeslee, J. P. 1999, AJ, 118, 1506

Block, D. L., Bertin, G., Stockton, A., Grosbol, P., Moorwood, A. F. M., & Peletier, R. F. 1994, A&A, 288, 365

Block, D., Freeman, K. C., Puerari, I., Combes, F., Buta, R., Jarrett, T., & Worthey, G. 2004a, in Penetrating Bars Through Masks of Cosmic Dust, eds.

D. L. Block, I. Puerari, K. C. Freeman, R. Groess, & E. Block (Kluwer: Springer), 15

Block, D. L., Freeman, K. C., Jarrett, T. H., Puerari, I., Worthy, G., Combes, F., & Groess, R. 2004b, A&A, 425, L37

Block, D. L., Puerari, I., Elmegreen, B. G., Elmegreen, D. M., Fazio, G. G., & Gehrz, R. D. 2009, ApJ, 694, 115

Book, L. G., & Benson, A. J. 2010, ApJ, 716, 810

Bothun, G., Impey, C. D., Malin, D. F., & Mould, J. R. 1987, AJ, 94, 23

Bournaud, F., & Combes, F. 2002, A&A, 392, 83

Brook, C., Governato, F., Quinn, T., Wadsley, J., Brooks, A. M., Willman, B., Stilp, A., & Jonsson, P. 2008, ApJ, 689, 678

Bureau, M., Aronica, G., Athanassoula, E., Dettmar, R.-J., Bosma, A., & Freeman, K. C. 2006, MNRAS, 370, 753

Bush, S. J., & Wilcots, E. M. 2004, AJ, 128, 2789

Buta, R. 1987, ApJS, 64, 1

Buta, R. 1995, ApJS, 96, 39

Buta, R., & Block, D. L. 2001, ApJ, 550, 243

Buta, R., & Combes, F. 1996, Fundam Cosm Phys, 17, 95

Buta R., & Crocker D. A. 1993, AJ, 106, 939

Buta, R., & Zhang, X, 2009, ApJS, 182, 559

Buta, R., Mitra, S., de Vaucouleurs, G., & Corwin, H. G. 1994, AJ, 107, 118

Buta, R., Byrd, G., & Freeman, T. 2003, AJ, 125, 634

Buta, R., Laurikainen, E., Salo, H., Block, D. L., & Knapen, J. H. 2006, AJ, 132, 1859

Buta, R. J., Corwin, H. G., & Odewahn, S. C. 2007, The de Vaucouleurs Atlas of Galaxies (Cambridge: Cambridge University Press (dVA))

Buta, R. J., Knapen, J. K., Elmegreen, B. G., Salo, H., Laurikainen, E., Elmegreen, D. M., Puerari, I., & Block, D. L. 2009, AJ, 137, 4487

Buta, R. et al. 2010a, ApJS, 190, 147

Buta, R., Laurikainen, E., Salo, H., & Knapen, J. H. 2010b, ApJ, 721, 259

Butcher, H., & Oemler, A. 1978, ApJ, 219, 18

Byrd, G. G., Rautiainen, P. Salo, H., Buta, R., & Crocker, D. A. 1994, AJ, 108, 476

Comerón, E. et al. 2010, MNRAS, 409, 354

Caon, N., Capaccioli, M., & D'Onofrio, M. 1993, MNRAS, 265, 1013

Cappellari, M. et al. 2011, MNRAS, 416, 1680

Cardemone, C. et al. 2009, MNRAS, 399, 1191

Carignan, C., & Purton, C. 1998, ApJ, 506, 125

Carollo, C. M., Stiavelli, M., de Zeeuw, P. T., & Mack, J. 1997, AJ, 114, 2366

Cayatte, V., Kotanyi, C., Balkowski, C., & van Gorkom, J. H. 1994, AJ, 107, 1003

Chapman, S. C., Windhorst, R., & Odewahn, S., et al. 2003, ApJ, 599, 92

Chung, A., van Gorkom, J. H., Kenney, J. D. P., Crowl, H., & Vollmer, B. 2009, AJ, 138, 1741

Cole, A. A. et al. 2007, ApJ, 659, L17

Comerón, S., Martinez-Valpuesta, I., Knapen, J. H., & Beckman, J. 2009, ApJ, 706, L25

Comerón, S., Knapen, J. H., Beckman, J. E., Laurikainen, E., Salo, H., Martinez-Valpuesta, I., & Buta, R. J. 2010, MNRAS, 402, 2462

Conselice, C. J. 1997, PASP, 109, 1251

Conselice, C. J. 2003, ApJS, 147, 1

Conselice, C. J. 2009, MNRAS, 399, L16

Conselice, C. J., & Arnold, J. 2009, MNRAS, 397, 208

Conselice, C. J. et al. 2004, ApJ, 600, L139

Conselice, C. J., Rajgor, S., & Myers, R. 2008, MNRAS, 386, 909

Contopoulos, G. 1996, ASP Conf Ser, 91, 454

Contopoulos, G., & Grosbol, P. 1989, A&A Rev, 1, 261

Cowie, L., Hu, E., & Songaila, A. 1995, AJ, 110, 1576

Crocker, D. A., Baugus, P. D., & Buta, R. 1996, ApJS, 105, 353

Cullen, H., Alexander, P., Green, D. A., & Sheth, K. 2007, MNRAS, 376, 98

Darg, D. W. et al. 2010, MNRAS, 401, 1043

Davies, R. L., Efstathiou, G., Fall, S. M., Illingworth, G., & Schechter, P. L. 1983, ApJ, 266, 41

de Vaucouleurs, G. 1956, Mem Comm Obs, 3(13), 1

de Vaucouleurs, G. 1958, ApJ, 127, 487

de Vaucouleurs, G. 1959, Handb Phys, 53, 275

de Vaucouleurs, G. 1961, ApJS, 5, 233

de Vaucouleurs, G. 1963, ApJS, 8, 31

de Vaucouleurs, G., & Freeman, K. C. 1972, Vistas Astron, 14, 163

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., Buta, R., Paturel, G., & Fouque, P. 1991, Third Reference Catalogue of Bright Galaxies (New York: Springer) (RC3)

de Vaucouleurs, G., de Vaucouleurs, A., & Corwin, H. G. 1976, Univ. Texas Mono. Astr. No. 2, Second Reference Catalogue of Bright Galaxies (Austin: University of Texas Press) (RC2)

de Zeeuw, P. T. et al. 2002, MNRAS, 329, 513

Djorgovski, S. 1992, in Morphological and Physical Classification of Galaxies, eds. G. Longo, M. Capaccioli, & G. Busarello (Dordrecht: Kluwer), 337

Dressler, A. 1980, ApJ, 236, 351

Durbala, A., Sulentic, J. W., Buta, R., & Verdes-Montenegro, L. 2008, MNRAS, 390, 881

Durbala, A., Buta, R., Sulentic, J. W., & Verdes-Montenegro, L. 2009, MNRAS, 397, 1756

Elmegreen, D. M. 1981, ApJS, 47, 229

Elmegreen, D. M., & Elmegreen, B. G. 1987, ApJ, 314, 3

Elmegreen, D., & Elmegreen, B. G. 2006, ApJ, 651, 676 (EE06)

Elmegreen, B. G., Elmegreen, D. M., & Montenegro, L. 1992, ApJS, 79, 37

Elmegreen, D., Elmegreen, B. G., & Sheets, C. 2004a, ApJ, 603, 74 (EES04)

Elmegreen, D., Elmegreen, B. G., & Hirst, A. 2004b, ApJ, 604, L21 (EEH04)

Elmegreen, D., Elmegreen, B. G., & Ferguson, T. E. 2005, ApJ, 623, L71

Elmegreen, D., Elmegreen, B. G., Ferguson, T. E., & Mullan, B. 2007a, ApJ, 663, 734 (EEFM07)

Elmegreen, D. M., Elmegreen, B. G., Ravindranath, S., & Coe, D. A. 2007b, ApJ, 658, 763

Elmegreen, B. G., Bournaud, F., & Elmegreen, D. M. 2008, ApJ, 688, 67

Elmegreen, B. G., Elmegreen, D. M., Fernandez, M. X., & Lemonias, J. J. 2009a, ApJ, 692, 12

Elmegreen, D. M., Elmegreen, B. G., Marcus, M. T., Shainyan, K., Yau, A., & Petersen, M. 2009b, ApJ, 701, 306

Emsellem, E. et al. 2007, MNRAS, 379, 401

Eskridge, P. B. et al. 2000, AJ, 119, 536

Eskridge, P. B. et al. 2002, ApJS, 143, 73

Fazio, G. G. et al. 2004, ApJS, 154, 10

Ferguson, H. C., & Binggeli, B. 1994, A&ARev, 6, 67

Ferrarese, L. et al. 2006, ApJS, 164, 334

Freeman, K. C. 1975, in Galaxies and the Universe, eds. A. Sandage, M. Sandage, & J. Kristian (Chicago: University of Chicago Press), 409

Garcia-Ruiz, I., Sancisi, R., & Kuijken, K. 2002, A&A, 394, 769

Giavalisco, M. et al. 2004, ApJ, 600, L93

Gil de Paz, A., Madore, B. F., & Pevunova, O. 2003, ApJS, 147, 29

Gil de Paz, A. et al. 2005, ApJ, 627, L29

Gillett, F. C., Forrest, W. J., & Merrill, K. M. 1973, ApJ, 183, 87

Giovanelli, R., & Haynes, M. P. 1985, AJ, 292, 404

Graham, A. W., & Guzmán, R. 2003, AJ, 125, 2936

Grouchy, R. D., Buta, R., Salo, H., Laurikainen, E., & Speltincx, T. 2008, AJ, 136, 980

Grouchy, R. D., Buta, R. J., Salo, H., & Laurikainen, E. 2010, AJ, 139, 2465

Gunn, J. E., & Gott, J. R. 1972, ApJ, 176, 1

Gunn, J. E. et al. 1998, AJ, 116, 3040

Helfer, T. T., Thornley, M. D., Regan, M. W., Wong, T., Sheth, K., Vogel, S. N., Blitz, L., & Bock, D. C.-J. 2003, ApJS, 145, 259

Helou, G. et al. 2004, ApJS, 154, 253

Higdon, J. L. 1995, ApJ, 455, 524

Higdon, J. L., Buta, R. J., & Purcell, G. B. 1998, AJ, 115, 80

Holmberg, E. 1950, Medd Lunds Astron Obs Ser II, 128

Holwerda, B. W., Keel, W. C., Williams, B., Dalcanton, J. J., & de Jong, R. S. 2009, AJ, 137, 3000

Hubble, E. 1926, ApJ, 64, 321

Hubble, E. 1936, The Realm of the Nebulae (Yale: Yale University Press)

Hubble, E. 1943, ApJ, 97, 112

Hunt, L. K., & Malkan, M. A. 1999, ApJ, 516, 660

Hunter, D. A. 1997, PASP, 109, 937

Impey, C., & Bothun, G. D. 1997, ARA&A, 35, 267

Irwin, M. J. et al. 2007, ApJ, 656, L13

Jarrett, T. H. et al. 2003, AJ, 125, 525

Jeans, J. 1929, Astronomy and Cosmogony (Cambridge: Cambridge University press)

Jerjen, H., Kalnajs, A., & Binggeli, B. 2000, A&A, 358, 845

Jogee, S. et al. 2009, ApJ, 697, 1971

Jokimaki, A., Orr, H., & Russell, D. G. 2008, AP&SS, 315, 249

Karachentseva, V. E. 1973, Astrofiz Issled-Izv Spets Astrofiz Obs, 8, 3

Karataeva, G. M., Tikhonov, N. A., Galazutdinova, O. A., Hagen-Thorn, V. A., & Yakovleva, V. A. 2004, A&A, 421, 833

Kennicutt, R. C., Tambln, P., & Congdon, C. W. 1994, ApJ, 435, 22

Kerr, F., & de Vaucouleurs, G. 1955, Aust J Phys, 8, 508

Knapen, J. H. 2005, A&A, 429, 141

Knapen, J. H. 2010, in Galaxies and Their Masks, eds. D. L. Block, K. C. Freeman, & I. Puerari (New York: Springer), in press

Knapen, J. H., & James, P. A. 2009, ApJ, 698, 1437

Knapen, J., Beckman, J. E., Shlosman, I., Peletier, R. F., Heller, C. H., & de Jong, R. S. 1995a, ApJ, 443, L73

Knapen, J., Beckman, J. E., Heller, C. H., Shlosman, I., & de Jong, R. S. 1995b, ApJ, 454, 623

Knapen, J. H., Shlosman, I., & Peletier, R. F. 2000, ApJ, 529, 93

Koopmann, R., & Kenney, J. D. P. 2004, ApJ, 613, 866

Kormendy, J. 1979, ApJ, 227, 714

Kormendy, J. 1985, ApJ, 295, 73

Kormendy, J. 1999, ASPC, 182, 124

Kormendy, J., & Bender, R. 1996, ApJ, 464, L119

Kormendy, J., & Bender, R. 2012, ApJS, 198, 2

Kormendy, J., & Djorgovski, S. 1989, ARA&A, 27, 235

Kormendy, J., & Kennicutt, R. C. 2004, ARA&A, 42, 603 (KK04)

Kormendy, J., & Norman, C. A. 1979, ApJ, 233, 539

Kormendy, J., Fisher, D. B., Cornell, M. E., & Bender, R. 2009, ApJS, 182, 216

Kormendy, J., Drory, N., Bender, R. A., & Cornell, M. E. 2010, ApJ, 723, 54

Kuijken, K., & Merrifield, M. R. 1995, ApJ, 443, L13

Laine, S., Shlosman, I., Knapen, J., & Peletier, R. F. 2002, ApJ, 567, 97

Laurikainen, E., & Salo, H. 2002, MNRAS, 337, 1118

Laurikainen, E., Salo, H., & Buta, R. 2004, ApJ, 607, 103

Laurikainen, E., Salo, H., & Buta, R. 2005, MNRAS, 362, 1319

Laurikainen, E., Salo, H., Buta, R., Knapen, J., Speltincx, T., & Block, D. L. 2006, MNRAS, 132, 2634

Laurikainen, E., Salo, H., Buta, R., & Knapen, J. 2007, MNRAS, 381, 401

Laurikainen, E., Salo, H., Buta, R., & Knapen, J. 2009, ApJ, 692, L34

Laurikainen, E., Salo, H., Buta, R., Knapen, J. H., & Comerón, S. 2010, MNRAS, 405, 1089

Laurikainen, E., Salo, H., Buta, R., & Knapen, J. 2011, MNRAS, 418, 1452

Ledo, H. R., Sarzi, M., Dotti, M., Khichfar, S., & Morelli, L. 2010, MNRAS, 407, 969

Lelli, F., Fraternali, F., & Sancisi, R. 2010, A&A, 516, 11

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. 2008, Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389, 1179–1189

Lintott, C. et al. 2009, MNRAS, 399, 129

Lisker, T., Debattista, V. P., Ferreras, I., & Erwin, P. 2006, MNRAS, 370, 477

Lotz, J. M., Madau, P., Giavalisco, M., Primack, J., & Ferguson, H. C. 2006, ApJ, 636, 592

Lupton, R., Blanton, M. R., Fekete, G., Hogg, D. W., O'Mullane, W., Szalay, A., & Wherry, N. 2004, PASP, 116, 133

Lynden-Bell, D., & Kalnajs, A. J. 1972, MNRAS, 157, 1

Madore, B. F., Nelson, E., & Petrillo, K. 2009, ApJS, 181, 572

Malin, D. F., & Carter, D. 1980, Nature, 285, 643

Malin, D. F., & Carter, D. 1983, ApJ, 274, 534

Martin, P., & Friedli, D. 1997, A&A, 326, 449

Martin, D. C. et al. 2005, ApJ, 619, L1

Martin, D. C. et al. 2007, ApJS, 173, 342

Martínez-Delgado, D., Peñarrubia, J., Gabany, R. J., Trujillo, I., & Majewski, S. 2008, ApJ, 689, 184

Martínez-Delgado, D., et al. 2010, AJ, 140, 962

Martinez-Valpuesta, I., Shlosman, I., & Heller, C. 2006, ApJ, 637, 214

Martinez-Valpuesta, I., Knapen, J. H., & Buta, R. 2007, AJ, 134, 1863

Masters, K. L. et al. 2010, MNRAS, 405, 783

Mateo, M. 1998, ARA&A, 36, 435

Matthews, T. A., Morgan, W. W., & Schmidt, M. 1964, ApJ, 140, 35

Mazzuca, L. M., Swaters, R. A., Veilleux, S., & Knapen, J. H. 2009, BAAS, 41, 693

McGaugh, S., Schombert, J. M., & Bothun, G. D. 1995, AJ, 109, 2019

McKernan, B., Ford, K. E. S., & Reynolds, C. S. 2010, astro-ph 1005.4907

Menanteau, F., Abraham, R. G., & Ellis, R. S. 2001, MNRAS, 322, 1

Menanteau, F. et al. 2004, ApJ, 612, 202

Meurer, G. R., Carignan, C., Beaulieu, S. F., & Freeman, K. C. 1996, AJ, 111, 1551

Michard, R., & Marchal, J. 1993, A&AS, 98, 29

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613

Morgan, W. W. 1958, PASP, 70, 364

Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427

Odewahn, S. C. 1991, AJ, 101, 829

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., & Zumach, W. A. 1992, AJ, 103, 318

Oosterloo, T. A., Morganti, R., Sadler, E. M., Vergani, D., & Caldwell, N. 2002, AJ, 123, 729

Pahre, M., Ashby, M. L. N., Fazio, G. G., & Willner, S. P. 2004, ApJS, 154, 235

Pence, W. D., Taylor, K., Freeman, K. C., de Vaucouleurs, G., & Atherton, P. 1988, ApJ, 326, 564

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2010, AJ, 139, 2097

Pereira-Santaella, M., Alonso-Herrero, A., Rieke, G. H., Colina, L. Díaz-Santos, T., Smith, J.-D. T., Pérez-González, P. G., & Engelbracht, C. W. 2010, ApJS, 188, 447

Quillen, A. C., Frogel, J. A., & González, R. A. 1994, ApJ, 437, 162

Rahman, N., Howell, J. H., Helou, G., Mazzarella, J. M., & Buckalew, B. 2007, ApJ, 663, 908

Rautiainen, P., & Salo, H. 2000, A&A, 362, 465

Rautiainen, P., Salo, H., & Buta, R. 2004, MNRAS, 349, 933

Ravindranath, S. et al. 2006, ApJ, 652, 963

Regan, M., & Teuben, P. 2003, ApJ, 582, 723

Regan, M., & Teuben, P. 2004, ApJ, 600, 595

Regan, M. W., Thornley, M. D., Helfer, T. T., Sheth, K., Wong, T., Vogel, S. N., Blitz, L., & Bock, D. C.-J. 2001, ApJ, 561, 218

Regan, M. et al. 2004, ApJS, 154, 204

Revaz, Y., & Pfenniger, D. 2007, in Astrophys. & Space Sci. Proc., Island Universes, ed. R. de Jong (New York: Springer), 149

Reynolds, J. H. 1927, Observatory, 50, 185

Rieke, G. H., & Low, F. J. 1972, ApJ, 176, L95

Rix, H.-W., & Rieke, M. J. 1993, ApJ, 418, 123

Rix, H.-W. et al. 2004, ApJS, 152, 163

Romano, R., Mayya, Y. D., & Vorobyov, E. I. 2008, AJ, 136, 1259

Romero-Gómez, M., Masdemont, J. J., Athanassoula, E., & García-Gómez, C. 2006, A&A, 453, 39

Romero-Gómez, M., Athanassoula, E., Masdemont, J. J., & García-Gómez, C. 2007, A&A, 472, 63

Ryder, S. D., Buta, R. J., Toledo, H., Shukla, H., Staveley-Smith, L., & Walsh, W. 1996, ApJ, 460, 665

Sackett, P. D., Rix, H.-W., Jarvis, B. J., & Freeman, K. C. 1994, ApJ, 436, 629

Saha, K., de Jong, R., & Holwerda, B. 2009, MNRAS, 396, 409

Salo, H., & Laurikainen, E. 2000a, MNRAS, 319, 377

Salo, H., & Laurikainen, E. 2000b, MNRAS, 319, 393

Salo, H., Laurikainen, E., Buta, R., & Knapen, J. H. 2010, ApJ, 715, L56

Sandage, A. 1961, Carnegie Inst. of Wash. Publ. No. 618, The Hubble Atlas of Galaxies (Washington: Carnegie Institution of Washington)

Sandage, A. 1975, in Stars and Stellar Systems, Vol. IX, Galaxies and the Universe, eds. A. Sandage, M. Sandage, & J. Kristian (Chicago: University of Chicago Press), 1

Sandage, A. 2004, in Penetrating Bars Through Masks of Cosmic Dust, ed. D. L. Block et al. (Dordrecht: Springer), 39

Sandage, A. 2005, ARA&A, 43, 581

Sandage, A., & Bedke, J. 1994, Carnegie Inst. of Wash. Pub. No. 638, The Carnegie Atlas of Galaxies (Washington: Carnegie Institution of Washington)

Sandage, A., & Binggeli, B. 1984, AJ, 89, 919

Sandage, A., & Tammann, G. A. 1981, Carnegie Inst. of Wash. Pub. No. 635, A Revised Shapley-Ames Catalog of Bright Galaxies (1st ed.; Washington: Carnegie Institution of Washington) (RSA)

Sanders, D. B., & Mirabel, I. F. 1996, ARA&A, 34, 749

Schawinski, K., Thomas, D., Sarzi, M., Maraston, C., Kavaraj, S., Joo, S., Yi, S. K., & Silk, J. 2007, MNRAS, 382, 1415

Schombert, J. 1986, ApJS, 60, 603

Schombert, J. 1987, ApJS, 64, 643

Schombert, J. 1988, ApJ, 328, 475

Schombert, J., Pildis, R. A., Eder, J. A., & Oemler, A. 1995, AJ, 110, 2067

Schwarz, M. P. 1981, ApJ, 247, 77

Schwarz, M. P. 1984, MNRAS, 209, 93

Schweizer, F. 1998, in Galaxies: Interactions and Induced Star Formation, ed. R. C. Kennicutt, et al. (Berlin: Springer), 105

Schweizer, F., & Seitzer, P. 1988, ApJ, 328, 88

Schweizer, F., Ford, W. K., Jedrzejewski, R., & Giovanelli, R. 1987, ApJ, 320, 454

Schweizer, F., van Gorkom, J., & Seitzer, P. 1989, ApJ, 338, 770

Scoville, N. et al. 2007, ApJS, 172, 38

Seiden, P. E., & Gerola, H. 1982, Fund Cosm Phys, 7, 241

Seigar, M. S., Block, D. L., Puerari, I., Chorney, N. E., & James, P. A. 2005, MNRAS, 359, 1065

Seigar, M. S., Kennefick, D., Kennefick, J., & Lacy, C. H. 2008, ApJ, 678, L93

Sellwood, J. A. 2010, in Planets, Stars, and Stellar Systems, Vol. 5, in production, ed. G. F. Gilmore (New York/Heidelberg: Springer)

Sellwood, J. A., & Wilkinson, A. 1993, Rep Prog Phys, 56, 173

Shamir, L. 2009, MNRAS, 399, 1367

Sheth, K. et al. 2008, ApJ, 675, 1141

Sheth, K. et al. 2010, PASP, 122, 1397

Sil'chenko, O. K., & Afanasiev, V. L. 2004, AJ, 127, 2641

Simkin, S. M., Su, H. J., & Schwarz, M. P. 1980, ApJ, 237, 404

Skrutskie, M. F. et al. 2006, AJ, 131, 1163

Sparke, L. S., van Moorsel, G., Erwin, P., & Wehner, E. M. H. 2008, AJ, 135, 99

Spitzer, L., & Baade, W. 1951, ApJ, 113, 413

Steidel, C., & Hamilton, D. 1992, AJ, 104, 941

Steidel, C., Giavalisco, M., Dickinson, M., & Adelberger, K. 1996, AJ, 112, 352

Steinmetz, M., & Navarro, J. F. 2002, NewA, 7, 155

Strateva, I. et al. 2001, AJ, 122, 1861

Struck, C. 2010, MNRAS, 403, 1516

Struck, C., Appleton, P. N., Borne, K. D., & Lucas, R. A. 1996, AJ, 112, 1868

Sulentic, J. et al. 2006, A&A, 449, 937

Surace, J. A., Sanders, D. B., Vacca, W. D., Veilleux, S., & Mazzarella, J. M. 1998, ApJ, 492, 116

Taniguchi, Y., & Watanabe, M. 1987, ApJ, 313, 89

Taylor-Mager, V. A., Conselice, C. J., Windhorst, R. A., & Jansen, R. A. 2007, ApJ, 659, 162

Theis, C., Sparke, L., & Gallagher, J. 2006, A&A, 446, 905

Theys, J. C., & Spiegel, J. C. 1976, ApJ, 208, 650

Thilker, D. A. et al. 2010, ApJ, 714, L171

Thornley, M. 1996, ApJ, 469, 45

Thuan, T. X., & Martin, G. E. 1981, ApJ, 247, 823

Thuan, T. X., Izotov, Y. I., & Lipovetsky, V. A. 1997, ApJ, 477, 661

Tolstoy, E., Hill, V., & Tosi, M. 2009, ARA&A, 47, 371

Treuthardt, P., Salo, H., Rautinainen, P., & Buta, R. 2008, AJ, 136, 300

Väisänen, P., Ryder, S., Mattila, S., & Kotilainen, J. 2008, ApJ, 689, L37

van den Bergh, S. 1976, ApJ, 206, 883

van den Bergh, S. 1980a, PASP, 92, 122

van den Bergh, S. 1980b, PASP, 92, 409

van den Bergh, S. 1995, AJ, 110, 613

van den Bergh, S. 1998, Galaxy Morphology and Classification (Cambridge: Cambridge University Press)

van den Bergh, S. 2009a, ApJ, 694, L120

van den Bergh, S. 2009b, ApJ, 702, 1502

van den Bergh, S., Pierce, M., & Tully, R. B. 1990, ApJ, 359, 4

van den Bergh, S., Abraham, R. G., Ellis, R. S., Tanvir, N. R., Santiago, B. X., & Glazebrook, K. G. 1996, AJ, 112, 359

van den Bergh, S., Cohen, J. G., Hogg, D. W., & Blandford, R. 2000, AJ, 120, 2190

van der Wel, A., Bell, E. F., Holden, B. P., Skibba, R. A., & Rix, H.-W. 2010, ApJ, 714, 1779

van Driel, W. et al. 1995, AJ, 109, 942

Vanzi, L., & Sauvage, M. 2004, A&A, 415, 509

Verdes-Montenegro, L., Sulentic, J., Lisenfeld, U., Leon, S., Espada, D., Garcia, E., Sabater, J., & Verley, S. 2005, A&A, 436, 443

Veron-Cetty, M., & Veron, P. 2006, A&A, 455, 773

Volonteri, M., Saracco, P., Chincarini, G., & Bolzonella, M. 2000, A&A, 362, 487

Walter, F., Brinks, E., de Blok, W. J. G., Bigiel, F., Kennicutt, R. C., Thornley, M. D., & Leroy, A. 2008, AJ, 136, 2563

White, R. E., & Keel, W. C. 1992, Nature, 359, 129

Whitmore, B. C., Lucas, R. A., McElroy, D. B., Steiman-Cameron, T. Y., Sackett, P. D., & Olling, R. P. 1990, AJ, 100, 1489

Williams, R. E. et al. 1996, ApJ, 112, 1335

Willman, B. et al. 2005, ApJ, 626, L85

Willner, S. et al. 2004, ApJS, 154, 222

Windhorst, R. et al. 2011, ApJS, 193, 27

Wirth, A., & Gallagher, J. S. 1984, ApJ, 282, 85

Wyder, T. K. et al. 2007, ApJS, 173, 293

Yagi, M., Yoshida, M., Komiyama, Y., Kashikawa, N., Furusawa, H., Okamura, S., Graham, A. W., Miller, N. A., Carter, D., Mobasher, B., & Jogee, S. 2010, AJ, 140, 1814

York, D. G. et al. 2000, AJ, 120, 1579

Zhang, X. 1996, ApJ, 457, 125

Zhang, X. 1998, ApJ, 499, 93

Zhang, X. 1999, ApJ, 518, 613

Zhang, X., & Buta, R. 2007, AJ, 133, 2584

2 Elliptical and Disk Galaxy Structure and Modern Scaling Laws

Alister W. Graham
Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Australia

1 Introduction 93
1.1 Early Beginnings 94
1.2 The Modern Galaxy 95

2 Elliptical Galaxy Light Profiles 97
2.1 Sérsic's Model 99
2.1.1 Dark Matter Halos 101
2.2 The Core-Sérsic Model 102
2.2.1 Central Mass Deficits 105
2.3 Excess Nuclear Light 105
2.4 Excess Halo Light 107

3 Structure-Related Scaling Relations 109
3.1 Linear Relations 110
3.1.1 Luminosity-(Central Surface Brightness) Relation 110
3.1.2 Luminosity-Concentration Relation 111
3.1.3 Concentration-(Central Surface Brightness) Relation 112
3.2 Curved Relations 112
3.2.1 Luminosity-(Effective Surface Brightness) Relation 112
3.2.2 Size-Luminosity Relation 113
3.2.3 Size-Concentration Relation 114
3.2.4 Size-(Effective Surface Brightness) Relation 115
3.3 Broken Relations 117
3.3.1 Luminosity-(Central Surface Brightness) Relation 117
3.3.2 Luminosity-Color Relation 119
3.3.3 Dynamics 119

4 Disk Galaxy Light Profiles 120
4.1 The Bulge-Disk Decomposition 121
4.2 Dust and Inclination Corrections 122
4.3 Pseudobulges 124
4.3.1 Sérsic Index 125
4.3.2 Rotation 126

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_2, © Springer Science+Business Media Dordrecht 2013

4.3.3 Ages 126
4.3.4 Scaling Relations 127
4.4 Bulgeless Galaxies 128
4.5 Barred Galaxies 129

5 Summary 130

Acknowledgments 131

References 131

Abstract: A century ago, in 1911 and 1913, Plummer and then Reynolds introduced their models to describe the radial distribution of stars in "nebulae." This article reviews the progress since then, providing both an historical perspective and a contemporary review of the stellar structure of bulges, disks, and elliptical galaxies. The quantification of galaxy nuclei, such as central mass deficits and excess nuclear light, plus, briefly, the structure of dark matter halos and cD galaxy envelopes, are discussed. Issues pertaining to spiral galaxies including dust, bulge-to-disk ratios, bulgeless galaxies, bars, and the identification of pseudobulges are also reviewed. An array of modern scaling relations involving sizes, luminosities, surface brightnesses, and stellar concentrations are presented, many of which are shown to be curved. These "redshift zero" relations not only quantify the behavior and nature of galaxies in the Universe today but are the modern benchmark for evolutionary studies of galaxies, whether based on observations, N-body simulations, or semi-analytical modeling. For example, it is shown that some of the recently discovered compact elliptical galaxies at $1.5 < z < 2.5$ may be the bulges of modern disk galaxies.

Keywords: Galaxy: Bars, Galaxy: Bulge-disk ratios, Galaxy: Bulgeless, Galaxy: Central mass deficits, Galaxy: cD, halos, Galaxy: Compact, Core-Sérsic model, Dark matter halos, Galaxy: Dust, Einasto model, Galaxy: Elliptical, Galaxy: Elliptical, Hubble, Edwin, Galaxy: Nuclei, Galaxy: Pseudobulges, Reynolds, Joseph B., Galaxy: Scaling relations, Sérsic model, Galaxy: Spiral, Galaxy: Structure

1 Introduction

For the last century, astronomers have been modeling the structure of "nebulae," and here, we focus on those external to the Milky Way. A key activity performed by many astronomers, past and present, is the categorization of these galaxies (Sandage 2005) and the quantification of their physical properties. How big are they? How bright are they? What characteristics distinguish or unite apparent subpopulations? Answers to such questions and the establishment of "scaling relations" between two or more galactic properties provide valuable insight into the physical mechanisms that have shaped galaxies.

Understanding how galaxies form, increasingly through the use of simulations and semi-analytic modeling (e.g., Kauffmann et al. 2004; Di Matteo et al. 2005; Croton et al. 2006; Naab et al. 2006; Nipoti et al. 2006; Covington et al. 2011; Guo et al. 2011; De Lucia et al. 2006; Bower et al. 2006, and references therein), requires an accurate knowledge of galaxy properties and scaling laws, as elucidated by Driver (2010). Not surprising, our knowledge of galaxies is best in the nearby Universe – out to distances, typically measured in megaparsecs rather than by redshift z, where galaxy structures can be reasonably well resolved. The properties of these galaxies provide the $z = 0$ benchmark used in the calibration of galactic evolutionary studies – both observed and simulated.

Popular scaling relations involving global galaxy parameters such as size, surface brightness, luminosity, and concentration are reviewed here. As we shall see, many bivariate distributions, which are frequently assumed to be linear, are often only approximately so over a restricted luminosity range. For example, the useful Kormendy (1977b) relation is only the tangent to the bright arm of a continuous but curved effective radius-(surface brightness) relation which unifies dwarf and giant elliptical galaxies (➋ Sect. 3.2.4). Similarly, the Faber and Jackson (1976) relation with a slope of 4 represents the average slope over a restricted luminosity range to a curved or broken luminosity-(velocity dispersion) distribution, in which the slope is 2 rather than 4 at lower luminosities (➲ Sect. 3.3.3). Knowing these trends, the bulk of which cannot be established when assuming structural homology, i.e., using de Vaucouleurs' (1948) $R^{1/4}$ model, is vital if one is to measure, model, and make sense of galaxies.

This article has been structured into four main sections. ➲ Section 1 provides this general overview plus a further review and introduction to galaxies on the Hubble-Jeans sequence.[1] Included are diagrams showing the location of dynamically hot stellar systems in the mass-size and mass-density plane, revealing that some high-z compact galaxies have properties equivalent to the bulges of local disk galaxies. ➲ Section 2 provides a historical account of how the radial distribution of stars in elliptical galaxies has been modeled and the iterative steps leading to the development of the modern core-Sérsic model (➲ Sect. 2.2). Subsections cover the Sérsic model (➲ Sect. 2.1), its relation and applicability to dark matter halos (➲ Sect. 2.1.1), partially depleted galaxy cores (➲ Sect. 2.2.1), excess nuclear light (➲ Sect. 2.3), and excess light at large radii in the form of halos or envelopes around giant elliptical galaxies (➲ Sect. 2.4). ➲ Section 3 presents and derives a number of elliptical galaxy scaling relations pertaining to the main body of the galaxy. From just two linear relations which unite the faint and bright elliptical galaxy population (➲ Sect. 3.1), a number of curved relations are derived (➲ Sect. 3.2). Several broken relations, at $M_B \approx -20.5$ mag, are additionally presented in ➲ Sect. 3.3. For those interested in a broader or different overview of elliptical galaxies, some recent good reviews include Renzini (2006), Cecil and Rose (2007), Ciotti (2009), and Lisker (2009). Finally, the latter third of this chapter is tied up in ➲ Sect. 4 which contains a discussion of the light profiles of disk galaxies and their bulge-disk decomposition (➲ Sect. 4.1). Also included are subsections pertaining to dust (➲ Sect. 4.2), the difficulties with identifying pseudobulges (➲ Sect. 4.3), potential bulgeless galaxies (➲ Sect. 4.4), and methods to model bars (➲ Sect. 4.5). Throughout the article, references to often overlooked discovery or pioneer papers are provided.

1.1 Early Beginnings

Looking out into the Milky Way arced across our night sky, the notion that we are residents within a pancake-shaped galaxy seems reasonable to embrace. Indeed, back in 1750, Thomas Wright also conjectured that we reside within a flat layer of stars which is gravitationally bound and rotating about some center of mass. However, analogous to the rings of Saturn, he entertained the idea that the Milky Way is comprised of a large annulus of stars rotating about a distant center, or that we are located in a large thin spherical shell rotating about some divine center (one of the galactic poles). While he had the global geometry wrong, he was perhaps the first to speculate that faint, extended nebulae in the heavens are distant galaxies with their own (divine) centers.

As elucidated by Hoskin (1970), it was Immanuel Kant (1755), aware of the elliptically shaped nebulae observed by Maupertuis and working from an incomplete summary of Wright (1750) that had been published in a Hamburg Journal,[2] who effectively introduced the modern concept of disklike galactic distributions of stars – mistakenly crediting Wright for the idea.

[1] This review does not encompass dwarf spheroidals, or any, galaxies fainter than $M_B \approx -14$. These galaxies cannot be observed (to date) at cosmologically interesting distances, and their increased scatter in the color-magnitude relation may indicate a range of galaxy types (e.g., Penny and Conselice 2008, and references therein).

[2] Freye Urtheile, Achtes Jahr (Hamburg 1751), translated by Hastie, op. cit., Appendix B.

Using his 1.83-m "Leviathan of Parsonstown" metal reflector telescope in Ireland, Lord William Henry Parsons, the third Earl of Rosse, discovered 226 New General Catalogue[3] (NGC: Dreyer 1888) and 7 Index Catalogue (IC: Dreyer 1895, 1908) objects (Parson 1878). Important among these was his detection of spiral structure in many galaxies, such as M51, which became known as the Whirlpool Galaxy.

Further divisions into disk (spiral) and elliptical galaxy types followed (e.g., Wolf 1908; Knox Shaw 1915; Curtis 1918; Reynolds 1920; Hubble 1926),[4] and Shapley and Swope (1924) and Shapley (1928) successfully identified our own Galaxy's (gravitational) center toward the constellation Sagittarius (see also Seares 1928).

With the discovery that our Universe contains redshifted "nebulae" that are expanding away from us (de Sitter 1917; Slipher 1917; see also Friedmann 1922; Lundmark 1924, and the review by Kragh and Smith 2003; Shaviv 2011), in accord with a redshift distance relation (Lemaitre 1927; Robertson 1928; Humasson 1929; Hubble 1929)[5] – i.e., awareness that some of the "nebulae" are external objects to our galaxy – came increased efforts to categorize and organize these different types of "galaxy." As noted by Sandage (2004, 2005), Sir James Jeans (1928) was the first to present the (tuning fork)-shaped diagram that encapsulated Hubble's (1926) early-to-late-type galaxy sequence, a sequence which had been inspired in part by Jeans (1919) and later popularized by Hubble (1936a; see Block et al. 2004). Quantifying the physical properties of galaxies along this sequence, with increasing accuracy and level of detail, has occupied many astronomers since. Indeed, this review addresses aspects related to the radial concentration of stars in the elliptical and disk galaxies which effectively define the Hubble-Jeans sequence. Irregular galaxies are not discussed here.

1.2 The Modern Galaxy

For reasons that will become apparent, this review uses the galaxy notation of Alan Sandage and Bruno Binggeli, in which dwarf elliptical (dE) galaxies are the faint extension of ordinary and luminous elliptical (E) galaxies, and the dwarf spheroidal (dSph) galaxies – prevalent in our Local Group (Grebel 2001) – are found at magnitudes fainter than $M_B \approx -13$ to -14 ($\approx 10^8 M_\odot$ in stellar mass; see ➋ *Fig. 2-1a*). ➋ *Figure 2-1a* reveals a second branch of elliptically shaped objects stretching from the bulges of disk galaxies and compact elliptical (cE) galaxies to ultracompact dwarf (UCD) objects (Hilker et al. 1999; Drinkwater et al. 2000; Norris and Kannappan 2011, and references therein). A *possible* connection is based upon the stripping of a disk galaxy's outer disk to form a cE galaxy (Nieto 1990; Bekki et al. 2001b; Graham 2002; Chilingarian et al. 2009) and through greater stripping of the bulge to form a UCD (Zinnecker et al. 1988; Freeman 1990; Bassino et al. 1994; Bekki et al. 2001a). It is thought that nucleated dwarf elliptical galaxies may also experience this stripping process, giving rise to UCDs.

While the identification of local spiral galaxies is relatively free from debate, the situation is not so clear in regard to elliptically shaped galaxies. The discovery of UCDs, which have

[3] The NGC built upon the (Herschel family's) Catalogue of Nebulae and Clusters of Stars (Herschel 1864).

[4] Reynolds (1927) called Hubble's attention to preexisting and partly similar galaxy classification schemes which were not cited.

[5] It is of interest to note that Hubble (1934, 1936b, 1937) was actually cautious to accept that the redshifts corresponded to real velocities and thus an expanding Universe as first suggested by others. He used the term "apparent velocity" to flag his skepticism. In point of fact, Hubble and Tolman (1935) wrote that the data were "not yet sufficient to permit a decision between recessional or other causes for the red-shift."

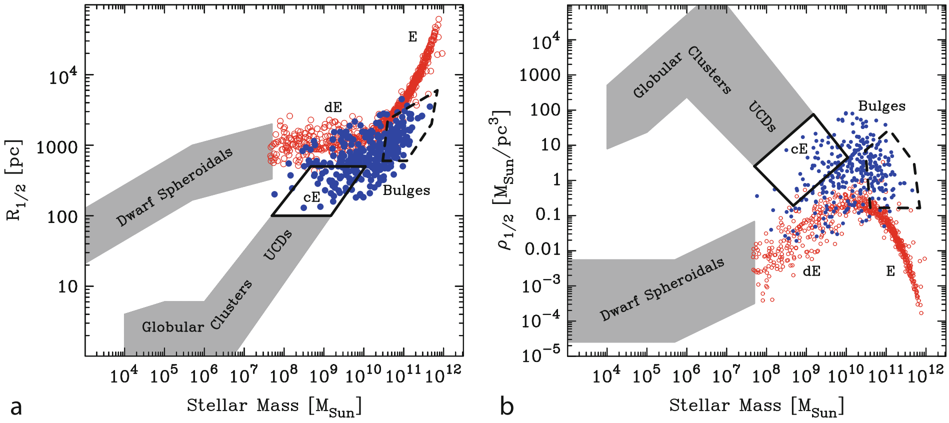


Fig. 2-1

(a) The radius containing half of each object's light, $R_{1/2}$ (as seen in projection on the sky), is plotted against each object's stellar mass. *Open circles*: Dwarf elliptical (*dE*) and ordinary elliptical (*E*) galaxies from Binggeli and Jerjen (1998), Caon et al. (1993), D'Onofrio et al. (1994), and Forbes et al. (2008). *Filled circles*: Bulges of disk galaxies from Graham and Worley (2008). *Shaded regions* adapted from Binggeli et al. (1984, their Fig. 7), Dabringhausen et al. (2008, their Fig. 2), Forbes et al. (2008, their Fig. 7), Misgeld and Hilker (2011, their Fig. 1). The location of the so-called "compact elliptical" (*cE*) galaxies is shown by the rhombus overlapping with small bulges. The location of dense, compact, $z = 1.5$ galaxies, as indicated by Damjanov et al. (2009, their Fig. 5), is denoted by the dashed boundary overlapping with luminous bulges. (b) Stellar mass density within the volume containing half each object's light, $\rho_{1/2}$, versus stellar mass. The radius of this volume was taken to equal 4/3 × $R_{1/2}$ (Ciotti 1991; Wolf et al. 2010)

sizes and fluxes intermediate between those of galaxies and (i) the nuclear star clusters found at the centers of galaxies and (ii) globular clusters (GCs, e.g., Haşegan et al. 2005; Brodie and Strader 2006), led Forbes and Kroupa (2011) to try and provide a modern definition for what a galaxy is (see also Tollerud et al. 2011). Only a few years ago, there was something of a divide between GCs and UCDs – all of which had sizes less than ~30 pc – and galaxies with sizes greater than 120 pc (Gilmore et al. 2007). However, as we have steadily increased our celestial inventory, objects of an intermediate nature have been found (e.g., Ma et al. 2007, their Table 3), raising the question asked by Forbes and Kroupa for which, perhaps not surprisingly, no clear answer has yet emerged. While those authors explored the notion of a division by, among other properties, size and luminosity, they did not discuss how the density varies. As an addendum of sorts to Forbes and Kroupa (2011), the density of elliptically shaped objects is presented here in ➋ *Fig. 2-1b*. This is also done to allow the author to wave the following flag.

Apparent in ➋ *Fig. 2-1b*, but apparently not well recognized within the community, is that the bulges of disk galaxies can be much denser than elliptical galaxies. If the common idea of galaxy growth via the accretion of a disk, perhaps from cold-mode accretion streams, around a pre-existing spheroid is correct (e.g., Navarro and Benz 1991; Steinmetz and Navarro 2002; Birnboim and Dekel 2003; see also Conselice et al. 2011 and Pichon et al. 2011), then one should expect to find dense spheroids at high z with 10^{10}–$10^{11} M_{\odot}$ of stellar material, possibly surrounded by a faint (exponential) disk which is under development. It is noted here that the dense, compact early-type galaxies recently found at redshifts of 1.4–2.5 (Daddi et al. 2005; Trujillo et al. 2006) display substantial overlap with the location of present day bulges in ➋ *Fig. 2-1a* and that the merger scenario to convert these compact high-z galaxies into today's elliptical galaxies is not without its problems (e.g., Nipoti et al. 2009; Nair et al. 2011). It is also noted that well-developed disks and disk galaxies are rare at the redshifts where these compact objects have been observed alongside normal-sized elliptical galaxies. Before trying to understand galaxy structure at high redshift, and galaxy evolution – themes not detailed in this review – it is important to first appreciate galaxy structures at $z = 0$, where observations are easier and local benchmark scaling relations have been established.

2 Elliptical Galaxy Light Profiles

Over the years, a number of mathematical functions have been used to represent the radial distribution of stellar light in elliptical galaxies, i.e., their light profiles. Before getting to de Vaucouleurs' $R^{1/4}$ model in the following paragraph, it seems apt to first quickly mention some early competitors. Although Plummer's (1911) internal-density model was developed for the nebulae which became known as globular clusters, because of its simplicity, it is still used today by some researchers to simulate elliptical galaxies, even though, it should be noted, no modern observers use this model to describe the radial distribution of light in elliptical galaxies. Reynolds' (1913) surface-density model, sometimes referred to as Hubble's (1930) model or the Reynold-Hubble model, was used to describe the nebula which became known as elliptical galaxies. It has an infinite mass and is also no longer used by observers today. The modified Hubble model (Rood et al. 1972), which also has an infinite mass, is also still sometimes used by simulators, even though, again, observers do not use this model anymore. Oemler's (1976) exponentially truncated Hubble model, known as the Oemler-Hubble model, is also not used to represent the observed stellar distribution in elliptical galaxies because it too, like its predecessors, was simply an approximation applicable over a limited radial range, as noted by King (1978). It is interesting to note that up until the 1980s, departures at large radii from the Reynolds model were attributed to tidal stripping by external gravitational potentials. That is, for three quarters of a century, Reynolds' model – originally developed from low-quality data for one galaxy – was generally thought to describe the original, undisturbed stellar distribution in elliptical galaxies.

de Vaucouleurs' (1948, 1953) $R^{1/4}$ surface-density model had traction for many years, in part due to de Vaucouleurs (1959) arguing that it fits better than the Reynolds model used by Hubble – a point reiterated by Kormendy (1977a) and others – and the revelation that it fits the radially extended data for NGC 3379 exceedingly well (de Vaucouleurs and Capaccioli 1979). Hodge (1961a, b) had, however, revealed that de Vaucouleurs' model was inadequate to describe faint elliptical galaxies, and Hodge (1963, 1964), in addition to King (1962),[6] noted that the 3-parameter King model, with its flatter inner profile and steeper decline at large radii, did a better job. For a time, King's (1962, 1966) model became popular for describing the light distribution in faint elliptical galaxies, at least until the exponential model – also used for the disks of spiral galaxies – was noted to provide a good description of some dwarf elliptical galaxies (Hodge 1971; Faber and Lin 1983; Binggeli et al. 1984) and that these galaxies need not have

[6] King (1962) also noted that his model failed to fit the inner region of bright elliptical galaxies.

experienced any tidal truncation (a prescription of the King model with its tidal radius parameter). Lauer (1984, 1985) additionally showed that King's modified isothermal model, with its flat inner core, was inadequate to describe the deconvolved light profiles of ordinary elliptical galaxies with "cores," i.e., galaxies whose inner light profile displays a nearly flat core. King's model does, however, remain extremely useful for studies of star clusters, globular clusters,[7] dwarf spheroidal galaxies, and galactic satellites which, unlike ordinary elliptical galaxies, can have flat cores in their inner surface brightness profile.

Today, the model of choice for describing nearby (and distant) dwarf and ordinary elliptical galaxies is Sérsic's (1963) generalization of de Vaucouleurs' $R^{1/4}$ model to give the $R^{1/n}$ surface-density model (➲ Sect. 2.1). This model reproduces the exponential model when $n = 1$ and de Vaucouleurs' model when $n = 4$; it can thus describe the main body of faint and luminous elliptical galaxies. The key advantage that this model has is (i) its ability to describe the observed stellar distributions that have a range of central concentrations (known to exist since at least Reaves 1956), and (ii) it provides a very good description of the data over (almost) the entire radial extent. Indeed, departures in the light profile from a well-fit Sérsic's model invariably signal the presence of additional features or components, rather than any failing of the model. Expanding upon the Sérsic model, the core-Sérsic model (➲ Sect. 2.2) is nowadays used to quantify those galaxies with "cores."

Although referring to the King model, the following quote from King (1966) seems particularly insightful "…de Vaucouleurs' law appears to refer to a particular central concentration and should be appropriate only for galaxy profiles that have that concentration." While noted by others, such as Oemler (1976), Capaccioli (1985), Michard (1985), and Schombert (1986), some three decades elapsed before the relevance of King's remark to elliptical galaxies resurfaced – albeit slowly at first – in the 1990s. Indeed, de Vaucouleurs' useful, albeit limited, $R^{1/4}$ model was referred to as a "law" for nearly half a century. However, we are now more keenly aware that (even normal) elliptical galaxies possess a range of central concentrations: concentrations which are well quantified in terms of the exponent n in Sérsic's $R^{1/n}$ model (see Trujillo et al. 2001; Graham et al. 2001). Use of $R^{1/4}$ model parameters alongside some model-independent measure of galaxy concentration is thus inconsistent, since every $R^{1/4}$ model has the same concentration.

Before introducing the equation for Sérsic's model in the following section, it is pointed out that in addition to modeling what can be considered the main body of the galaxy, one can also find excess stellar light at (i) small radii in the form of nuclear (i.e., centrally located) disks and dense nuclear star clusters (➲ Sect. 2.3) and also at (ii) large radii in the form of halos or envelopes in cD and central cluster galaxies (➲ Sect. 2.4). As briefly noted above, deficits of stellar flux at the cores of massive galaxies are also observed, and a model to quantify these stellar distribution, relative to the outer nondepleted light profile, is described in ➲ Sect. 2.2. While nonsymmetrical components in elliptical galaxies can also exist, they are not addressed here given the focus on well-structured systems. Somewhat random, non-symmetrical components may be a sign of a disturbed morphology, of ongoing nonuniform star formation (See ➲ Chap. 3) or gravitationally induced tidal features from external forces.

[7] The Wilson (1975) and Elson et al. (1987) model are also useful for describing globular clusters.

2.1 Sérsic's Model

José Sérsic's (1963, 1968) $R^{1/n}$ model, which was introduced in Spanish, describes how the projected surface-intensity I varies with the projected radius R, such that

$$I(R) = I_e \exp\left\{-b_n\left[\left(\frac{R}{R_e}\right)^{1/n} - 1\right]\right\} \tag{2.1}$$

and I_e is the intensity at the "effective" radius R_e that encloses half of the total light from the model (Ciotti 1991; Caon et al. 1993). The term b_n ($\approx 1.9992n - 0.3271$ for $0.5 < n < 10$, Capaccioli 1989) is not a separate parameter but is instead dependent on the third model parameter, n, that describes the shape, i.e., the concentration, of the light profile.[8] The exact value of b_n is obtained by solving the equation $\Gamma(2n) = 2\gamma(2n, b_n)$, where γ(2n,x) is the incomplete gamma function and Γ is the (complete) gamma function (Ciotti 1991). Useful Sérsic-related expressions have been presented in Ciotti (1991), Simonneau and Prada (2004), and Ciotti and Bertin (1999), while Graham and Driver (2005) provide a detailed review of Sérsic's model plus associated quantities and references to pioneers of this model.

The relation between the effective surface brightness ($\mu_e = -2.5 \log I_e$) and the central surface brightness ($\mu_0 = -2.5 \log I_0$) is given by the expression

$$\mu_e = \mu_0 + 1.086b, \tag{2.2}$$

where we have dropped the subscript n from the term b_n for simplicity, while

$$\langle\mu\rangle_e = \mu_e - 2.5 \log[e^b n\Gamma(2n)/b^{2n}] \tag{2.3}$$

gives the difference between the effective surface brightness and the mean effective surface brightness ($\langle\mu\rangle_e$) within R_e. *Figure 2-2* shows the behavior of the Sérsic model.

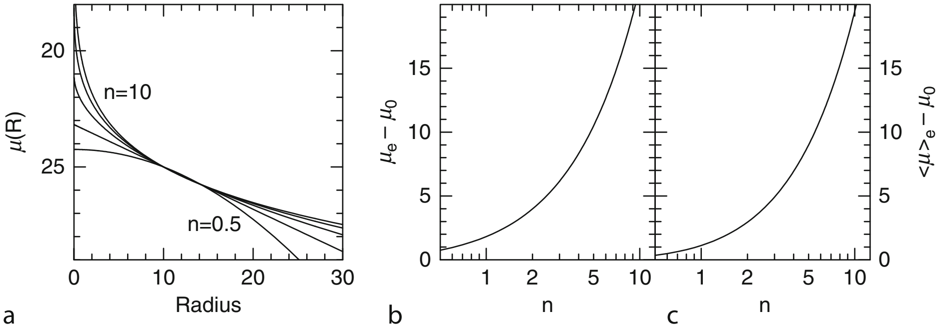


Fig. 2-2
***Left panel*: Sérsic's $R^{1/n}$ model (2.1) for indices n = 0.5, 1, 2, 4 and 10. The effective radius (R_e) and surface brightness ($\mu_e = -2.5 \log I_e$) have been arbitrarily set to 10 and 25. *Right panel*: Difference between the various surface brightness terms discussed in the text**

[8] Ellipticity gradients result in a different Sérsic index for the major- and minor-axis, as noted by Caon et al. (1993) and later quantified by Ferrari et al. (2004).

Uniting CCD data with wide and deep photographic images, Caon et al. (1990, 1993, 1994) revealed that the Sérsic $R^{1/n}$ model provided a remarkably good description to the stellar distribution over a large radial range, down to surface brightnesses of ~28 B-mag arcsec^{-2}, for the early-type galaxies brighter than $M_B = -18$ in the Virgo cluster. This work was in essence an expansion of de Vaucouleurs and Capaccioli's (1979) study of NGC 3379 which is very well fit with $n = 4$. Different galaxies were discovered by Caon et al. (1993) to be equally well fit but required different values of n (see also Bertin et al. 2002).

Importantly, Caon et al. (1993) additionally showed that a correlation existed between stellar concentration, i.e., the Sérsic index n, and (model independent) galaxy size that was not due to parameter coupling in the Sérsic model (see also Trujillo et al. 2001, their Sect. 2). One implications of this result is that $R^{1/4}$, and similarly Petrosian (1976), magnitudes, sizes, and surface brightnesses are systematically in error as a function of galaxy concentration (Graham et al. 2005; Hill et al. 2011). That is, application of a model which fails to adequately capture the range of stellar distributions will result in parameters which are systematically biased as a function of galaxy mass. For example, fitting an $R^{1/4}$ model to elliptical galaxies which are actually described by an $R^{1/n}$ model with n less than and greater than 4 will yield sizes and luminosities which are, respectively, greater than and less than the true value (e.g., Binggeli and Cameron 1991; Trujillo et al. 2001; Brown et al. 2003; Liu et al. 2008). Similarly, fitting an exponential model to bulges that are best described by an $R^{1/n}$ model with n less than and greater than 1 will yield sizes and luminosities which are, respectively, greater than and less than the true value (e.g., Graham 2001). Obviously, one does not want to fine tune galaxy simulations to match scaling relations that contain systematic biases due to poor measurements, and observers are therefore busy fitting $R^{1/n}$ models these days.

A good approximation to the internal-density profile associated with Sérsic's model, i.e., with its deprojection, was introduced by Prugniel and Simien (1997). Useful expressions for the dynamics, gravitational potential, and forces of this model have been developed by Trujillo et al. (2002), Terzić and Graham (2005), and Terzić and Sprague (2007). Somewhat more complex than the early light-profile models, such expressions can, importantly, accommodate a range of concentrations, rather than only varying one scale radius and one scale density. Such a model is vital if one wishes to properly simulate and understand the mass spectrum of elliptical galaxies, whose Sérsic index n increases with stellar mass. As stressed by Graham and Driver (2005), Sérsic's model and the core-Sérsic model 0 (➋ Sect. 2.2) have become key in unifying and understanding the galaxies around us.

Like the majority of surface- and internal-density models from the last century, the Sérsic function is an empirical model created to match data rather than developed from theory, and as such, we should be cautious before calling it a law. Attempts to find a physical explanation for de Vaucouleurs' model yielded results which helped to keep it in vogue. Dissipational models have long been touted for producing $R^{1/4}$ profiles (e.g., Larson 1969, 1974), and in the 1980s, papers based on dissipationless N-body simulations of a cold clumpy collapse or the merger of disk galaxies also claimed to finally produce $R^{1/4}$ (and also Reynolds) profiles (e.g., van Albada 1982; McGlynn 1984; Carlberg et al. 1986; Barnes 1988). However, a closer inspection reveals clear departures from the $R^{1/4}$ profile, with the simulated profiles better described by an $R^{1/n}$ model with $n < 4$. Obviously, their inability (or perhaps lack of desire (although see Farouki et al. 1983) whose nonhomologous merger remnants were initially criticized by $R^{1/4}$ aficionados) to create the range of stellar concentrations now observed in elliptical galaxies highlights a limitation of these early works. These pioneering studies led to N-body simulations by Nipoti et al. (2006) and Aceves et al. (2006) – and Farouki et al. (1983), whose results with a smaller force softening appeared years ahead of their time – which recover a range of Sérsic profile shapes for gravitational collapses in a dark matter halo and for disk galaxy mergers, respectively.

Given the empirical nature of Sérsic's $R^{1/n}$ model, Hjorth and Madsen (1995) revealed how dissipationless merging and violent relaxation provided a physical explanation for the departure from the homologous $R^{1/4}$ model. Other works have explained how the quasi-constant-specific entropy associated with the post violent-relaxation stage of elliptical galaxies results in the observed mass-dependent range of stellar concentrations in elliptical galaxies (Gerbal et al. 1997; Lima Neto et al. 1999; Márquez et al. 2001).

Elliptical galaxies, whether built by near-monolithic collapse, collisions of disk galaxies, and wet or dry mergers, appear to eventually experience the same force(s) of nature that results in their radial stellar distribution depending on the total stellar mass. That is, it may not matter how the mass was accumulated into an elliptical galaxy; once it becomes a dynamically heated, bound stellar system, it appears to eventually obey certain universal scaling relations (see ➜ Sect. 3).

It is interesting to note that Sérsic actually introduced his model as a way to parameterize disk galaxies which he thought were comprised of differing ratios of a disk plus an R$^{1/4}$ bulge. His model was not initially intended to fit elliptical galaxies, and as such it did not immediately threaten de Vaucouleurs' model. Credit for popularizing the use of Sérsic's $R^{1/n}$ model for approximating not only lenticular[9] bulge + disk galaxies but for describing pure elliptical galaxies resides largely with Massimo Capaccioli (e.g., Capaccioli 1985, 1987, 1989; Caon et al. 1993; D'Onofrio et al. 1994).[10] However, Davies et al. (1988) had also introduced this model for dwarf elliptical galaxies, while Sparks (1988) developed an early Gaussian seeing correction for this model and Ciotti (1991) developed a number of associated expressions such as the velocity dispersion profile and a distribution function. The important quantification that Capaccioli and others provided is how the radial distribution of stars in elliptical galaxies, i.e., their concentration, varies with the size, luminosity, and thus the mass of the elliptical galaxy (see also Cellone et al. 1994; Vennik and Richter 1994; Young and Currie 1994, 1995; Graham et al. 1996; Karachentseva et al. 1996; Vennik et al. 1996). As we shall see in this article, the implications of this breakthrough have been dramatic, unifying what had previously been considered two distinct species of galaxy, namely, dwarf and ordinary elliptical galaxies, previously thought to be described by an exponential and $R^{1/4}$ model, respectively.

2.1.1 Dark Matter Halos

This section would be somewhat incomplete without a few words regarding the connection between Sérsic's model and (simulated) dark matter halos. While modified theories of gravity may yet make dark matter redundant at some level, it is intriguing to note that the Prugniel and Simien (1997) internal-density model, developed to approximate the deprojected form of Sérsic's $R^{1/n}$ model, additionally provides a very good representation of the internal-density profiles of simulated dark matter halos. Merritt et al. (2006) found that it actually provides a better

[9] This term was introduced by Knox Shaw (1915) and Reynolds (1920) in their galaxy classification scheme.

[10] It is worth noting that D'Onofrio (2001) remodeled the Virgo and Fornax two-component lenticular galaxies with an $R^{1/n}$ bulge plus an exponential disk (see ➜ Sect. 4).

description than not only the Navarro et al. (1997) model but even a generalized NFW model with an arbitrary inner profile slope γ.

Sérsic's former student, Navarrro, independently applied Sérsic's surface-density model to the internal-density profiles of simulated dark matter halos (Navarro et al. 2004; Merritt et al. 2005). Jaan Einasto (1965) had previously developed this same function as Sérsic to describe the internal-density profiles of galaxies. Rather than a universal profile shape, as advocated by Navarro et al. (1997), a range of simulated dark matter density profile shapes is now known to vary with the dark matter halo mass (Avila-Reese et al. 1999; Jing and Suto 2000; Merritt et al. 2005; Del Popolo 2010, and references therein). A number of useful expressions related to this "Einasto model," which has the same functional form as Sérsic's model but is applied to the internal rather than projected density profile, can be found in Cardone et al. (2005), Mamon and Łokas (2005), and Graham et al. (2006).

An apparent "bulge-halo conspiracy" between the radial distribution of stellar mass and dark matter (after modification by baryons) has arisen in recent years, such that elliptical galaxies reportedly have *total* internal-density profiles $\rho(r)$ described by power laws (Bertin and Stiavelli 1993; Kochanek 1995; Gavazzi et al. 2007; Buote and Humphrey 2011). These power laws were originally claimed to be close to isothermal, such that $\rho(r) \propto r^{-2}$ (Koopmans et al. 2006, 2009; Gavazzi et al. 2007). Recent work emphasizes that only the sample average profile slope is close to −2, and that a trend in slope with galaxy size exists (Humphrey and Buote 2010; Auger et al. 2010). This is a developing field, and it is worth noting that the analyses have been confined to massive galaxies; dispersions greater than ~175 km s^{-1} and thus with Sérsic indices $n \gtrsim 4$ (Graham et al. 2001). While the light profile shape changes dramatically as the Sérsic index n increases from ~1 to ~4, there is not such an obvious change in light profile shape from ~4 to higher values of n (see ➲ *Fig. 2-2*). The apparent isothermal profiles of elliptical galaxies and the alleged "bulge-halo conspiracy" may turn out to be a by-product of sample selection (i.e., choosing galaxies which have approximately the same structure). It would be interesting to expand the Sloan Lens ACS (SLACS) Survey (Bolton et al. 2006) to a greater range than only bright early-type galaxies that are approximately well fit with an $R^{1/4}$ model and to go beyond the use of simple power-laws to describe the total mass density profile once the data allows this. Two-component mass models (e.g., Prugniel-Simien plus Einasto or Burkert (1995) or Persic et al. (1996), etc.) over a range of galaxy masses await. Claims that "early-type galaxies are structurally close to homologous" may therefore be premature, as was the case for the distribution of stellar light in elliptical galaxies while the $R^{1/4}$ model was thought to be a law.

Finally, although only mentioned here in passing, it is interesting to note that a number of recent works have announced the existence of a constant surface-density core for real dark matter halos of ~ 150 solar masses per square parsec (e.g., Kormendy and Freeman 2004; Spano et al. 2008; Donato et al. 2009).

2.2 The Core-Sérsic Model

The centers of luminous galaxies have long been known to possess "cores," such that the surface-density profile flattens at the center (e.g., King and Minkowski 1966), and King and Minkowski (1972) remarked on the inability of the Reynold's and de Vaucouleurs' model to match these flattened cores in giant galaxies. Although King (1978) identified a number of galaxies thought to be well described by his model, using seeing-deconvolved ground-based images, Lauer (1983, 1984, 1985) analyzed 14 galaxies with "cores," ranging from 1.5 to 5.0 arcseconds in radius, revealing that they, like M87 (Young et al. 1978; Duncan and Wheeler 1980; Binney and Mamon 1982), were not exactly described by the King model which had a completely flat core. Similar conclusions, that cores existed but that they do not have flat inner surface brightness profiles, were also reported by Kormendy (1982, 1985a), creating the need for a new model to describe the stellar distribution in galaxies.

Nearly a decade later, the Hubble Space Telescope was flying and offered factors of a few improvement over the best image resolution achievable from the ground at that time. Not surprisingly, astronomers explored the centers of galaxies. In an effort to quantify these nuclear regions, after the abandonment of the King model and the lack of a flattened core in the $R^{1/4}$ model, Crane et al. (1993), Ferrarese et al. (1994), Forbes et al. (1994), and Jaffe et al. (1994) used a double-power-law model to describe the inner light profiles of large galaxies. Grillmair et al. (1994), Kormendy et al. (1994), and Lauer et al. (1995) also adopted a double power-law model but one with an additional, fifth, parameter to control the sharpness of the transition. Their model, which they dubbed the "Nuker law" for describing the nuclear regions of galaxies (after excluding any apparent excess light), has the same functional form as the double power-law model presented by Hernquist (1990, his (43)) to describe the internal density of galaxies (Zhao 1996).

However, as noted by the above authors, these double power-law models were never intended to describe the entire radial extent of a galaxy's stellar distribution, and they provided no connection with the outer ($\mathrm{R}^{1/4}$ like) radial profile. This disconnection turned out to be their downfall. Due to the curved nature of the outer light profiles beyond the core, which were being fitted by the double power-law model's outer power law, the five parameters of the Nuker model systematically changed as the fitted radial extent changed. This was first illustrated in a number of diagrams by Graham et al. (2003) who revealed that none of the Nuker model parameters were robust, and as such they could not provide meaningful physical quantities. For example, Trujillo et al. (2004) reported that the Nuker-derived core-radii were typically double and up to a factor of 5 times larger than the radius where the inner power-law core broke away from the outer $R^{1/4}$-like profile – a result reiterated by Dullo and Graham (2012). In addition, these "break radii" were being identified in the so-called "power-law" galaxies that showed no evidence of a downward departure and flattening from the inward extrapolation of the outer $R^{1/4}$-like profile. This situation arose because of the curved nature of what were actually Sérsic profiles. That is, the so-called "power-law" galaxies not only had no distinct "core" like the "core galaxies" do, but confusingly, they do not even have power-law light profiles. Recent analysis using Nuker profiles has replaced the fitting parameters with a characteristic scale radius (r_γ in Lauer et al. 2007) which is defined nonparametrically by the profile slope. This makes the measurement robust, although now the mathematical nature of r_γ is quite different from previous break radii.

Given that Caon et al. (1993) and D'Onofrio et al. (1994) had established that the Sérsic function fits the brightness profiles of elliptical galaxies remarkably well over a large dynamic range (see the figures in Bertin et al. 2002), it is possible to confidently identify departures from these profiles that are diagnostic of galaxy formation. While in this and the following section we deal with partially depleted cores – also referred to as "missing light" – in luminous galaxies (thought to be built from dissipationless mergers, e.g., Khochfar and Burkert 2003), the ensuing section addresses extra central light above the inward extrapolation of the outer Sérsic profile (found in galaxies that have experienced dissipation and star formation).

Building on the work of Caon et al. (1993) and Graham et al. (2003) introduced the core-Sérsic model, which was applied in Trujillo et al. (2004). The model represents a modification to Sérsic's model such that it has an inner power-law core. In contrast to the Nuker team's (e.g., Kormendy et al. 1994; Lauer et al. 1995, 2005; Faber et al. 1997) combination of Nuker model parameters for the core with $R^{1/4}$ model parameters for the main galaxy, Graham et al. (2003), Graham and Guzmán (2003), Balcells et al. (2003), Graham (2004) and Trujillo et al. (2004) advocated the measurement of core properties, excesses and deficits of light, measured relative to the outer Sérsic model.

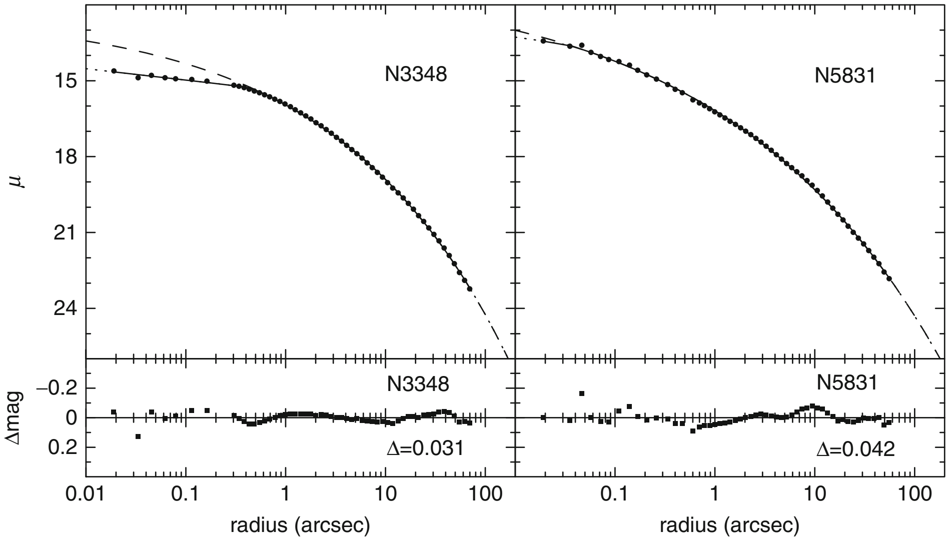
The core-Sérsic model provided an expression capable of unifying the nuclear regions of galaxies with their outer regions and also providing stable physical quantities.[11] The model can be written as

$$I(R) = I'\left[1+\left(\frac{R_b}{R}\right)^{\alpha}\right]^{\gamma/\alpha} \times \exp\left\{-b[(R^{\alpha}+R_b^{\alpha})/R_e^{\alpha}]^{1/(\alpha n)}\right\} \tag{2.4}$$

where R_b denotes the break radius separating the inner power-law having logarithmic slope γ from the outer Sérsic function. The intensity I_b at the break radius is such that

$$I' = I_b 2^{-(\gamma/\alpha)} \exp\left[b(2^{1/\alpha}R_b/R_e)^{1/n}\right]. \tag{2.5}$$

The sixth and final parameter, α, controls the sharpness of the transition between the inner (power-law) and the outer (Sérsic) regime – higher values of α indicating a sharper transition. ➲ *Figure 2-3* shows the core-Sérsic model (with $\alpha = 100$) applied to NGC 3348.



◘ Fig. 2-3
***Left*: Core-Sérsic model (*solid line*) fit to the major-axis *R*-band light profile of NGC 3348 (*dots*), with the *dashed line* showing the associated Sérsic component. The inner depleted zone corresponds to a stellar mass deficit of $3 \times 10^8 M_{\odot}$ (see Graham 2004). *Right*: NGC 5831 has, in contrast, no (obvious) partially depleted core and is well described by the Sérsic's model alone. The rms scatter is shown in the *lower panels* (Figure taken from Graham et al. (2003))**

[11] The core-Sérsic, and also Sérsic, model provides robust parameters beyond the core if one has sufficient data to sample the curvature in the light profile; in practice, this requires radial data out to ~$1R_e$.

Terzić and Graham (2005) and Terzić and Sprague (2007) provide a number of expressions related to the potential, force, and dynamics.[12] The core-Sérsic model is further discussed and used by Ferrarese et al. (2006a, b), Côté et al. (2006, 2007), Kawata et al. (2007), and Ciotti (2009).

2.2.1 Central Mass Deficits

The collisional construction of galaxies from the merger of lesser galaxies is thought to be a common occurrence in the Universe. Coupled with the presence of a supermassive black hole (SMBH) at the heart of most galaxies (Wolfe and Burbidge 1970; Magorrian et al. 1998), dissipationless mergers were proposed by Begelman et al. (1980; see also Ebisuzaki et al. 1991) to explain the depleted nuclei, i.e., the cores, observed in giant elliptical galaxies (e.g., King 1978 and references therein). It is thought that core depletion is primarily due to the gravitational slingshot (Saslaw et al. 1974) effect that coalescing binary SMBHs – from the pre-merged galaxies – have on stars while they themselves sink to the bottom of the potential well of the newly wed galaxy.[13]

Theory predicts that the central mass deficit M_{def} should scale with $0.5N\,M_{\mathrm{bh}}$, where M_{bh} is the final (merged) black hole mass and N the number of major "dry" (i.e., gas free, dissipationless) mergers (Milosavljević and Merritt 2001; Merritt and Milosavljević 2005; Merritt 2006a, b). Graham (2004) used the core-Sérsic model to quantify the central deficit of stars relative to the inward extrapolation of the outer Sérsic profile. ➲ *Figure 2-4* suggests that the luminous elliptical galaxies sampled have experienced an average of 1 or 2 major dry (i.e., dissipationless) mergers, a conclusion in agreement with select ΛCDM models of galaxy formation (Haehnelt and Kauffmann 2002; Volonteri et al. 2003).

Quantification of the central stellar deficit relative to the inward extrapolation of the outer Sérsic profile has also been applied to bright Virgo cluster galaxies by Ferrarese et al. (2006a), and Côté et al. (2007) and with the exception of VCC 798 – a lenticular (bulge plus disk) galaxy – provides similar results. Of course, when an outer disk exists, a core-Sérsic bulge plus disk fit is required, otherwise the disk will bias the Sérsic parameters.[14]

2.3 Excess Nuclear Light

While galaxies brighter than $M_B \sim -20.5$ have partially depleted cores (e.g., Faber et al. 1997; Rest et al. 2001; Graham and Guzmán 2003), fainter galaxies often contain additional stellar

[12] The complex nature of this model has resulted in the appearance of alternate expressions, e.g., Spergel (2010, see his Fig. 3) and Baes and Van Hese (2011).

[13] The presence of $FUV - NUV$ color gradients in core galaxies such as NGC 1399 suggests that they may not have been built from major, dry merger events (Carter et al. 2011). Additionally, the globular-cluster-specific frequency in core galaxies may be at odds with core-galaxy formation through equal-mass merger events because the fainter, intermediate luminosity elliptical galaxies may have lower specific frequencies (Harris and van den Bergh 1981). Alternative ideas for core formation in giant galaxies have been proposed: Boylan-Kolchin et al. (2004), Nipoti et al. (2006) and Martizzi et al. (2011).

[14] Trujillo et al. (2004) and Graham (2004) intentionally excluded lenticular galaxies from their publication because it was felt that the community were not yet ready for an 8-parameter model (six core-Sérsic plus two exponential-disk parameters). Intriguingly, VCC 798 resides near the 1-to-1 line in ➲ *Fig. 2-4* when an exponential-disk plus (core-Sérsic) bulge fit is performed (Graham, unpublished work).

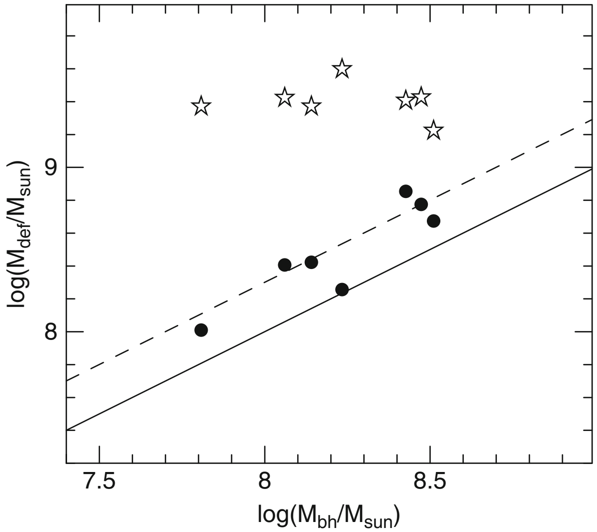


Fig. 2-4
Central mass deficit for seven core galaxies derived using the core-Sérsic model (*circles*) and the Nuker model (*stars*) plotted against each galaxy's predicted central supermassive black hole mass. The *solid* and *dashed line* shows M_{def} equal 1 and 2 M_{bh}, respectively (Figure adapted from Graham (2004))

components at their centers. Such excess nuclear light, above that of the underlying host galaxy, has long been known to exist (e.g., Smith 1935; Reaves 1956, 1977; Romanishin et al. 1977) and was systematically studied by Binggeli et al. (1984), van den Bergh (1986), and Binggeli and Cameron (1991, 1993) in a number of dwarf elliptical galaxies. As far back as Larson (1975), it was known that simulations containing gas can account for these dense nuclear star clusters, clusters which became easier to detect with the Hubble Space Telescope (e.g., Carollo et al. 1998).

For many years, it was common practice to simply exclude these additional nuclear components from the analysis of the galaxy light profile (e.g., Lauer et al. 1995; Byun et al. 1996; Rest et al. 2001; Ravindranath et al. 2001). Using HST data, Graham and Guzmán (2003) simultaneously modeled the host galaxy and the additional nuclear component with the combination of a Sérsic function for the host galaxy plus a Gaussian function for the nuclear star cluster. As we will see in Sect. 4, they also showed that the lenticular galaxies in their early-type galaxy sample could be modeled via a Sérsic-bulge plus an exponential-disk plus a Gaussian-(star cluster) decomposition of their light profile – as done by Wadadekar et al. (1999) with ground-based images. Many other studies have since modeled the nuclear star clusters seen in HST images, see *Fig. 2-5*, with the combination of a nuclear component plus a Sérsic host galaxy (e.g., Grant et al. 2005; Côté et al. 2006; Ferrarese et al. 2006a; Graham and Spitler 2009).

While Graham and Guzmán (2003) and Côté et al. (2006) found that some nuclear star clusters could actually be resolved, it is not yet established what mathematical function best describes their radial distribution of stars. The closest example we have to study is of course the 30 million solar mass nuclear star cluster at the center of the Milky Way (Launhardt et al. 2002). Graham and Spitler (2009) provided an analysis after allowing for the significant contamination

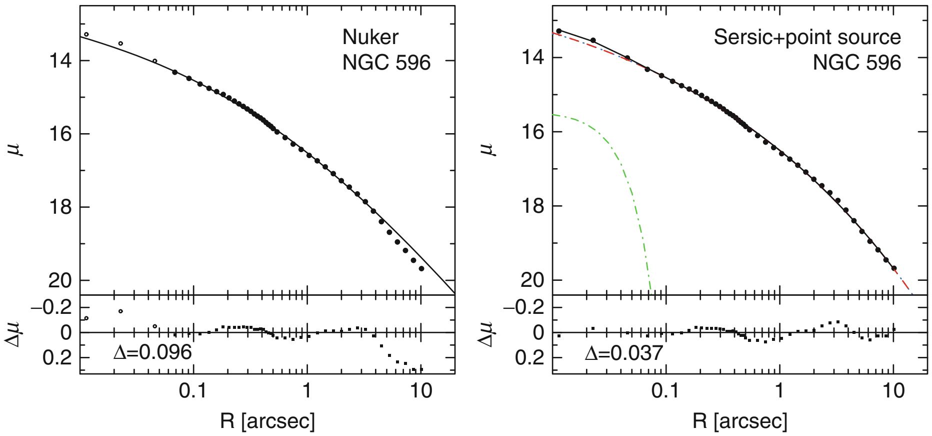


Fig. 2-5
***Left*: 5-parameter Nuker model fit to the *V*-band light profile of the nucleated galaxy NGC 596 after excluding the inner three data points (Lauer et al. 2005). *Right*: 3-parameter Sérsic model plus 2-parameter point source (Gaussian) fit to the same light profile of NGC 596. With the same number of parameters, this model fits both the inner, intermediate, and outer light profile (Figure from Dullo and Graham (2012))**

from bulge stars. Although they found that the cluster could be well described by a Sérsic index with $n = 3$ (see *Fig. 2-6*, it remains to be tested how well a King model can describe the data or at least the old stellar population known to have a core (e.g., Genzel et al. 1996). The excess nuclear light in M32 has also been well fit with an $n = 2.3$ Sérsic function (Graham and Spitler), but this too may yet be better described by a King model. What is apparent is that neither the underlying bulge nor the nuclear component is described by a power law. Theories which form or assume such power-law cusps appear to be at odds with current observations.

2.4 Excess Halo Light

This section discusses starlight from extended galaxy halos as it is reflected in intensity profiles; see Chap. 6 for a broader discussion of this component in the context of the evolution of galaxies in dense cluster environments.

Brightest cluster galaxies (BCGs), residing close to or at the centers of large galaxy clusters, have long been recognized as different from less luminous elliptical galaxies: their light profiles appear to have excess flux at large radii (e.g., Oemler 1976; Carter 1977; van den Bergh 1977; Lugger 1984; Schombert et al. 1986). However, before exploring this phenomenon, it is important to recall that the light profiles of large galaxies have Sérsic indices n greater than 4 (e.g., Caon et al. 1993; Graham et al. 1996).[15] Subsequently, at large radii in big elliptical galaxies, there will be excess flux above that of an $R^{1/4}$ model fit to some limited inner radial range.

An initially puzzling result from the Sloan Digital Sky Survey (*SDSS*; York et al. 2000) was the lack of light profiles with Sérsic $n > 5$–6 (Blanton et al. 2005a). This was however soon resolved

[15] As $n \to \infty$, the Sérsic model can be approximated by a power law (see Graham and Driver 2005).

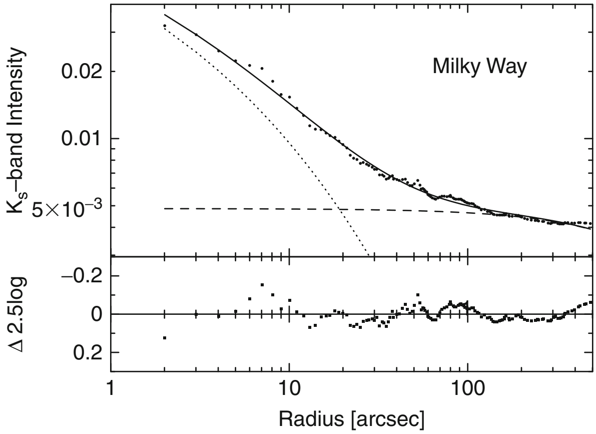


Fig. 2-6
Uncalibrated, 2MASS, K_s-band intensity profile from the center of the Milky Way, taken from Schödel et al. (2009, their Fig. 2). The nuclear star cluster is modeled with a Sérsic function with n = 3 (*dotted curve*) and the underlying host bulge with an exponential function (Kent et al. 1991) that has an effective half-light radius of ∼4.5 degrees (e.g., Graham and Driver 2007) and is therefore basically a *horizontal line*. One parsec equals 25 arcsec (Figure from Graham and Spitler (2009))

when Blanton et al. (2005b), Mandelbaum et al. (2005), and Lisker (2005, 2006b, 2007) pointed out a serious sky-subtraction problem with the early SDSS Photometric Pipeline (*photo*: Ivezić et al. 2004). The sky value to be subtracted from each galaxy had been measured too close to the galaxy in question, and because galaxies with large Sérsic indices possess rather extended light profiles, their outer galaxy light was actually measured and then subtracted from these galaxies. As a result, the high Sérsic index light profiles were erroneously erased and missed from the SDSS. This resulted in the $R^{1/4}$ model appearing to provide good fits to bright elliptical galaxies.

Bearing in mind that large elliptical galaxies have high Sérsic indices, it is important to distinguish between (i) an inadequacy of the $R^{1/4}$ model to describe what is actually a single $R^{1/n}$ profile and (ii) a distinct physical component such as an envelope of diffuse halo light surrounding a central galaxy. Early quantitative photometry of cD galaxies (supergiant D galaxies, e.g., Matthews et al. 1964; Morgan and Lesh 1965) suggested the presence of an inner $R^{1/4}$ spheroid plus an outer exponential corona (de Vaucouleurs 1969; de Vaucouleurs and de Vaucouleurs 1970). One should however question if this outer corona is a distinct entity or not. To answer this in the affirmative, astronomers can point to how the light profiles can display inflections marking the transition from BCG light to intracluster light. Gonzalez et al. (2005) additionally showed that the inflection in the light profiles of many BCGs was also associated with a change in the ellipticity profile, signaling the switch from BCG light to intracluster light.

Gonzalez et al. (2005) and Zibetti et al. (2005) chose to model their BCG sample using an $R^{1/4} + R^{1/4}$ model to describe the inner galaxy plus the outer halo light. However, given that elliptical galaxies are better described by the $R^{1/n}$ model, and the desire to measure the actual concentration of halos rather than assign a fixed $R^{1/4}$ profile, Seigar et al. (2007) fitted an $R^{1/n} + R^{1/n}$ model to the light profiles of five BCGs. An $R^{1/n}$ galaxy plus an exponential-halo model was found to provide the optimal fit in three instances, with an additional galaxy having no halo detected. The associated galaxy-to-halo luminosity ratios can be found there. This exponential, rather than $R^{1/4}$, behavior of the halo has been confirmed by Pierini et al. (2008). Intriguing is that the halo does not trace the NFW-like dark-matter halo density profiles produced in ΛCDM simulations (➋ Sect. 2.1.1). Stellar halos around non-BCG galaxies have also now been reported to display an exponential radial distribution (e.g., Gadotti 2011; Tal and van Dokkum 2011).

3 Structure-Related Scaling Relations

While it is common practice, and somewhat helpful, to call elliptical galaxies fainter than about $M_B = -18$ by the term "dwarf elliptical" (Sandage and Binggeli 1984), it should be noted, and this section will reveal, that on all measures, they appear to be the low-mass end of a continuous sequence which unifies dwarf and normal/luminous elliptical galaxies. Not only is this true in terms of their structural properties (e.g., Binggeli et al. 1984; Graham and Guzmán 2003; Gavazzi et al. 2005; Ferrarese et al. 2006a; Côté et al. 2006, 2007, 2008; Misgeld et al. 2008, 2009; Janz and Lisker 2009; Graham 2010; Chen et al. 2010; Glass et al. 2011; Ryś and Falcón-Barroso 2012) but even the degree of kinematically distinct components is similar (Chilingarian 2009).

There are many important relations between stellar luminosity, color, metallicity, age, and dynamics that have already revealed a continuous and linear behavior uniting dwarf and giant elliptical galaxies (e.g., Caldwell 1983; Davies et al. 1983; Binggeli et al. 1984; Bothun et al. 1986; Geha et al. 2003; Lisker and Han 2008). However, Kormendy et al. (2009, their Sect. 8) argue that these apparently unifying relations must not be sensitive to different physical processes, which they believe produce a dichotomy between dwarf and ordinary elliptical galaxies. They claim that it is only relations which show an apparent different behavior at the faint and bright end that are sensitive to the formation physics, and the remainder are not relevant. This section explains why such nonlinear relations are actually a consequence of the linear relations, and as such, these nonlinear relations actually support a continuum between dwarf and ordinary elliptical galaxies.

To begin, it should be reiterated that (dwarf and ordinary) elliptical galaxies – and the bulges of disk galaxies (➋ Sect. 4.1) – do not have structural homology. Instead, they have a continuous range of stellar concentrations – quantified by the Sérsic index n (Davies et al. 1988; Caon et al. 1993; D'Onofrio et al. 1994; Young and Currie 1994, 1995; Andredakis et al. 1995) – that varies linearly with both stellar mass and central surface brightness (after correcting for central deficits or excess light). A frequently unappreciated consequence of these two linear relations is that relations involving either the effective half-light radius (R_e) or the effective surface brightness (μ_e), or the mean surface brightness within R_e ($\langle\mu\rangle_e$), will be nonlinear. Such curved relations have often been heralded as evidence that two different physical processes must be operating because the relation is not linear and has a different slope at either end. To further complicate matters, sample selection which includes faint and bright elliptical galaxies, but excludes the intermediate-luminosity population, can effectively break such continuously curved relations into two apparently disconnected relations, as can selective colour-coding.

There are three distinct types of (two parameter) relations involving the properties of elliptical galaxies: (i) linear relations which are taken to reveal the unified nature of dEs and Es,

(ii) curved relations revealing a continuity that had in the past been misinterpreted to indicate that distinct formation process must be operating, and (iii) broken relations which do in fact imply that two physical mechanisms are operating. In the following sections, we shall learn how the linear relations result in the existence of curved relations when using effective radii and surface brightnesses. We shall also see that the transition in the broken relations occurs at $M_B \approx -20.5$ and thus has nothing to do with the previously held belief that dEs and Es are two distinct species separated at $M_B = -18$ (Wirth and Gallagher 1984; Kormendy 1985b, 2009).

3.1 Linear Relations

This section introduces two key relations, from which a third can be derived, involving structural parameters. They are the luminosity-concentration (L–n) relation and the luminosity-(central density)[16] (L–μ_0) relation. We shall use the Sérsic shape parameter n to quantify the central concentration of the radial light profile and use the projected central surface brightness μ_0 as a measure of the central density.[17]

It is noted that one can expect a certain level of scatter in the L–n and L–μ_0 diagrams because both the central density and the radial concentration of stars that one observes depends upon the chance orientation of one's triaxial galaxy (Binney 1978). This is of course also true for measurements of effective surface brightness, half-light radii, velocity dispersions, etc. To have the best relations, it is important that we use Sérsic parameters from elliptical galaxies rather than parameters from single-Sérsic fits to samples of elliptical and lenticular galaxies given the two-component (2D disk plus 3D bulge) nature of the lenticular galaxies. Given the offset nature of bulges and elliptical galaxies in the L–n diagram (e.g., Graham 2001; see also Möllenhoff and Heidt 2001) it is also important that bulges not be combined in this section's analysis of elliptical galaxy scaling relations.

3.1.1 Luminosity-(Central Surface Brightness) Relation

Caldwell (1983; his Fig. 6) and Bothun et al. (1986, their Fig. 7) revealed that, fainter than $M_B \approx -20.5$, there is a continuous linear relation between luminosity and central surface brightness. Furthermore, Binggeli et al. (1984, their Fig. 11) and Binggeli and Cameron (1991, their Figs. 9 and 18) revealed that, when using the inward extrapolation of King models, the L–μ_0 relation is continuous and roughly linear from $-12 > M_B > -23$. This same general result was also highlighted by Jerjen and Binggeli (1997) and Graham and Guzmán (2003) when using the inward extrapolation of the outer Sérsic model. The benefit of this approach is that one's central surface brightness is not biased by the presence of a depleted core or any additional nuclear components within the host galaxy. ➋ *Figure 2-7a* displays the elliptical galaxy (M_B, μ_0) data set from Graham and Guzmán (2003) fit by the expression

$$M_B = 0.67\mu_{0,B} - 29.5. \tag{2.6}$$

[16] Here the "central density" refers to the density prior to core depletion in giant elliptical galaxies or the growth of additional nuclear components in smaller elliptical galaxies.

[17] To convert from the surface density to the internal density, one can use (4) from Terzić and Graham (2005).

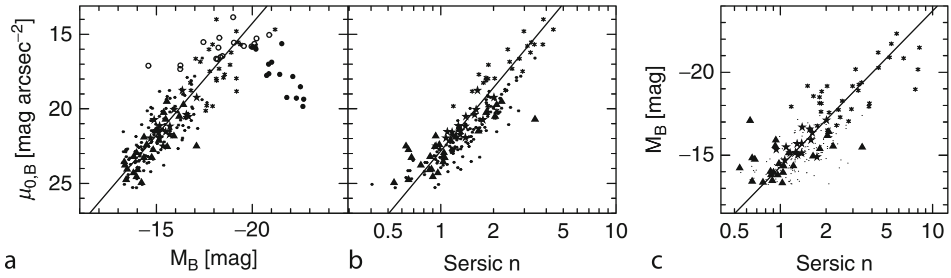


Fig. 2-7
Linear relations between the observed B-band central surface brightness ($\mu_{0,B}$) with (a) the absolute B-band magnitude M_B and (b) the Sérsic index n for a sample of elliptical galaxies. Panel (c) shows the galaxy magnitudes versus the Sérsic indices, with the line given by (2.7). The "core galaxies" (*large filled circles*) with partially depleted cores can be seen to have lower central surface brightnesses than the relation in panel (a). Inward extrapolation of their outer Sérsic profile yields μ_0 values which follow the linear relation, as previously noted by Jerjen and Binggeli (1997). The data have come from the compilation by Graham and Guzmán (2003, their Fig. 9). *Dots* represent dE galaxies from Binggeli and Jerjen (1998), *triangles* represent dE galaxies from Stiavelli et al. (2001), *large stars* represent Graham and Guzmán (2003) Coma dE galaxies, *asterisks* represent intermediate to bright E galaxies from Caon et al. (1993) and D'Onofrio et al. (1994), *open circles* represent the so-called "power-law" E galaxies from Faber et al. (1997), and the *filled circles* represent the "core" E galaxies from these same Authors

The actual central surface brightness of the luminous "core galaxies" is shown in *Fig. 2-7a*, rather than the value obtained from the inward extrapolation of their outer Sérsic profile. As such, these "core galaxies" were excluded from the fit, but see the discussion in Sect. 3.3. As an aside, if the central supermassive black hole mass M_{bh} in elliptical galaxies is directly related to the central stellar density (see Graham and Driver 2007), then the connections between M_{bh} and the global galaxy properties, such as total mass and velocity dispersion, may be secondary.

3.1.2 Luminosity-Concentration Relation

The linear relation between luminosity and Sérsic index, or strictly speaking the logarithm of these quantities, has been shown many times (e.g., Young and Currie 1994; Jerjen and Binggeli 1997; Graham and Guzmán 2003; Ferrarese et al. 2006a). This continuous relation between magnitude and concentration[18] for elliptical galaxies had of course been recognized before (e.g., Ichikawa et al. 1986, their Fig. 11). The following M_B–n expression is shown in *Fig. 2-7c*, again matched to the sample of elliptical galaxies compiled by Graham and Guzmán (2003).

$$M_B = -9.4\log(n) - 14.3. \tag{2.7}$$

[18] Graham et al. (2001) contains a review of various concentration indices used over the decades, while Trujillo et al. (2001) was the first to quantify the monotonic relation between Sérsic index and concentration.

Graham and Guzmán (2003) excluded two-component lenticular galaxies fit by others with a single-component Sérsic model. It may be prudent to continue to exclude these galaxies even after a Sérsic-bulge plus exponential-disk fit because the M_B–n relation defined by bulges, at least in spiral galaxies, is different from that defined by elliptical galaxies (Graham 2001, his Fig. 14).

3.1.3 Concentration-(Central Surface Brightness) Relation

Combining the above two equations provides an expression between central surface brightness and Sérsic index such that

$$\mu_0 = 22.8 - 14.1\log(n), \tag{2.8}$$

which is shown in ➲ *Fig. 2-7b*, where it can be seen to be roughly applicable for values of $n \gtrsim 1$.

3.2 Curved Relations

As a direct result of the above linear relations – which unite dwarf and giant elliptical galaxies – expressions involving either the effective half-light radius R_e, the associated effective surface brightness μ_e at this radius, or the mean surface brightness $\langle\mu\rangle_e$ enclosed within this radius, are curved.

3.2.1 Luminosity-(Effective Surface Brightness) Relation

The following analysis is from Graham and Guzmán (2003).

Given the empirical M_B–n relation (➲ 2.7), one knows what the expected value of n is for some value of M_B. One can then convert the empirical M_B–μ_0 relation (➲ 2.6) into an M_B–μ_e relation using the exact relation between μ_0 and μ_e which depends only on the value of n (➲ 2.2). Doing so, one obtains the expression

$$\begin{aligned}\mu_e &= 1.5M_B + 44.25 + 1.086b,\\ &= 1.5[M_B + 14.3] + 22.8 + 1.086b,\end{aligned} \tag{2.9}$$

where $b \approx 1.9992n - 0.3271$ and (➲ 2.7) is used to replace n in terms of M_B, such that $n = 10^{-(14.3+M_B)/9.4}$.

One can similarly convert the empirical M_B–μ_0 relation (➲ 2.6) into an M_B–$\langle\mu\rangle_e$ relation using the exact relation between μ_0 and $\langle\mu\rangle_e$ which also depends only on the value of n (➲ 2.3). Doing this, one obtains the expression

$$\langle\mu\rangle_e = 1.5M_B + 44.25 + 1.086b - 2.5\log\left[\frac{e^b n\Gamma(2n)}{b^{2n}}\right], \tag{2.10}$$

where again $b \approx 1.9992n - 0.3271$ and (➲ 2.7) is used to replace n in terms of M_B. These curves can be seen in ➲ *Fig. 2-8* (adapted from Graham and Guzmán 2004).

Binggeli et al. (1984, their Fig. 8) and Capaccioli and Caon (1991) previously showed with empirical data that the M_B–$\langle\mu\rangle_e$ relation is curved. What was new in Graham and Guzmán (2003) was the explanation. In the past, evidence of non-linear relations involving parameters from dwarf and giant elliptical galaxies were heralded as evidence of a dichotomy. Galaxy

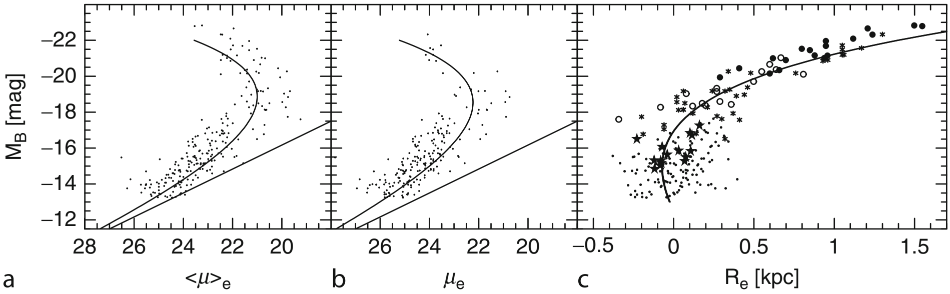


Fig. 2-8

Elliptical galaxy B-band magnitude versus (a) mean effective surface brightness (2.10), (b) effective surface brightness (2.9), and (c) effective radius (2.11). The continuous, curved relations are predictions from the observed linear relations in *Fig. 2-7*. The slight mismatch seen here arises from the rough fitting by eye of the linear relations (The data points have come from the compilation by Graham and Guzmán (2003, their Fig. 9))

sample selection that excluded the intermediate population, and therefore resulted in two apparently disconnected relations, acted to further convince some that they were dealing with two classes of object.[19]

3.2.2 Size-Luminosity Relation

Now that we know how to play this game, one can additionally make predictions for relations involving the effective radius R_e because we know that the luminosity $L = 2\pi\langle I\rangle_e R_e^2$, with $\langle\mu\rangle_e = -2.5\log\langle I\rangle_e$. As explained in Graham and Worley (2008, their Sect. 5.3.1), one can derive the size-luminosity relation such that

$$\log R_e[\mathrm{kpc}] = \frac{M_B}{10} + 1.066 + 0.434n + 0.5\log\left[\frac{b^{2n}}{\mathrm{e}^b n\Gamma(2n)}\right] \tag{2.11}$$

for $0.5 < n < 10$, with $b \approx 1.9992n - 0.3271$ and where (2.7) is used to replace n in terms of M_B. This size-luminosity relation for elliptical galaxies is shown in *Fig. 2-8c* along with real galaxy data.

Binggeli et al. (1984, their Fig. 7; cf. Misgeld and Hilker 2011, their Fig. 7) also demonstrated, with empirical data that the $L - R_e$ relation for dwarf and giant elliptical galaxies is curved. Their diagram, in addition to *Fig. 2-8c* seen here, reveals why studies which only sample bright elliptical galaxies are often content to simply fit a straight line (e.g., Kormendy 1977b). The explanation for why this happens is of course akin to approximating the Earth as flat when one (forgivably) does not sample enough of what is actually a curved profile. As Graham et al. (2006, their Fig. 1b) re-revealed recently, and as reiterated by Bernardi et al. (2007), a sample of massive elliptical galaxies will have a steeper size-luminosity relation than a sample of ordinary

[19] A simple one-dimensional example of sample bias would be a survey of the average physical properties, such as size or mass, of people at a primary school. One would measure the properties of children and adults but miss the bridging population which reveals a continuity and thus unification of the species.

elliptical galaxies, which will in turn have a steeper size-luminosity relation than a sample of dwarf elliptical galaxies because the size-luminosity relation is curved. Graham and Worley (2008) explains why the $L - R_e$ relation given by (➤ 2.11), based on two linear relations and the functional form of Sérsic's model, is curved.

Due to the linear relations between magnitude, central surface brightness, and the logarithm of the Sérsic exponent n, the use of faint isophotal radii results in what is roughly a linear size-luminosity relation (e.g., Oemler 1976; Strom and Strom 1978; Forbes et al. 2008, their Fig. 3; van den Bergh 2008; Nair et al. 2011), with the bright-end slope dependent on the adopted isophotal limit or Petrosian radius used. The implications of this important observation shall be detailed elsewhere.

Helping to propagate the belief that dwarf and ordinary elliptical galaxies are distinct species, Dabringhausen et al. (2008) and Lisker (2009) fit a double power law to their curved size-luminosity relation for dwarf and ordinary elliptical galaxies, thus yielding distinct slopes at the faint and bright end. In addition, the interesting study by Janz and Lisker (2008) reported small deviations from the predicted curved relation. However, galaxies that are well described by Sérsic's function and which follow linear M–n and M–μ_0 relations *must* follow a single-curved M–R_e relation. The deviations that they found are therefore mirroring (a) the inadequacy of the fitted linear relations to the M–n and M–μ_0 distribution (noted in the caption of ➤ *Fig. 2-8*) and/or (b) poorly fitting Sérsic models to their sample of elliptical *and* disk galaxies. Adding uncertainties to the linear relations in ➤ Sect. 3.1, and propagating those through to the predicted M–R_e relation, is required before we can claim evidence of significant deviations. However, searching for such second-order effects may indeed be interesting given the different types of dwarf galaxies that are emerging (Lisker et al. 2007).

3.2.3 Size-Concentration Relation

One can additionally derive an expression relating R_e and n by substituting the magnitude from the empirical M_B–n relation, expressed in terms of n (➤ 2.7), into the size-luminosity relation (➤ 2.11) to give

$$\begin{aligned} \log R_e[\mathrm{kpc}] &= 0.434n - 0.364 - 0.94\log(n) \\ &\quad +0.5\log\left[\frac{b^{2n}}{\mathrm{e}^b n\Gamma(2n)}\right]. \end{aligned} \tag{2.12}$$

While the M_B–n relation is linear, the R_e–n relation is curved, as can be seen in ➤ *Fig. 2-9*.

In passing, it is noted that the form of this relation (➤ 2.12) matches the bulge data from Fisher and Drory (2010, their Fig. 13). They interpret the departure of the low-n bulges ($n < 2$) from the approximately linear relation defined by the high-n bulges ($n > 2$) to indicate that a different formation process is operating to produce the less concentrated "pseudobulges." However, based upon linear unifying relations that span the artificial $n = 2$ divide, we know that this R_e–n relation must be curved. Without an understanding of this relation, and other curved relations (e.g., Greene et al. 2008), they have at times been misinterpreted and used to claim the existence of different physical processes (see ➤ Sect. 4.3 for a discussion of pseudobulges, and Hohl (1975) and references therein).

It may be worth better defining the behavior of the R_e–n relation at small sizes in ➤ *Fig. 2-9*. The data from Davies et al. (1988) suggests than when $n = 0.5$, values of R_e may range from 1 kpc down to 0.2 kpc (Caon et al. 1993, their Fig. 5). Such a reduction to the flattening of the

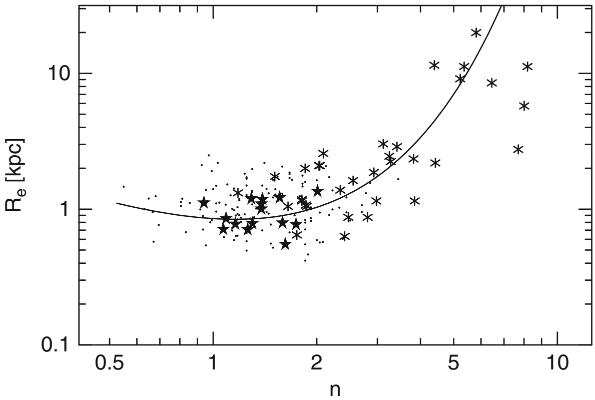


Fig. 2-9
Equation 2.12 is overplotted empirical data. *Symbols* have the same meaning as in *Fig. 2-7*

R_e–n distribution, below $n \approx 1$, may in part arise from the inclusion of dwarf spheroidal galaxies (see Misgeld and Hilker 2011, their Fig. 1).

3.2.4 Size-(Effective Surface Brightness) Relation

As discussed in Graham (2010), the first two linear relations in *Fig. 2-7* naturally explain the curved $\langle\mu\rangle_e$–R_e relation in *Fig. 2-10*. From the empirical μ_0–n relation ((2.8), *Fig. 2-7b*), one can convert μ_0 into $\langle\mu\rangle_e$ using (2.2) and (2.3). The effective radius R_e is acquired by matching the empirical M_B–μ_0 relation ((2.6), *Fig. 2-7a*) with the absolute magnitude formula

$$M = \langle\mu\rangle_e - 2.5\log(2\pi R^2_{e,\mathrm{kpc}}) - 36.57, \tag{2.13}$$

(see Graham and Driver 2005, their (12)). Eliminating the absolute magnitude gives the expression

$$\log R_e = \frac{1}{5}\left\{\frac{\langle\mu\rangle_e}{3} - 9.07 + 0.72b - 1.67\log\left(\frac{ne^b\Gamma(2n)}{b^{2n}}\right)\right\}, \tag{2.14}$$

in which we already know the value of n associated with each value of $\langle\mu\rangle_e$. This is achieved by (again) using the empirical μ_0–n relation (2.8) with (2.2) and (2.3), such that

$$\langle\mu\rangle_e = 22.8 + 1.086b - 14.1\log(n) - 2.5\log\left[\frac{ne^b\Gamma(2n)}{b^{2n}}\right] \tag{2.15}$$

and $b \approx 1.9992n - 0.3271$. Equation 2.14, obtained from two linear relations involving Sérsic parameters, is a curved relation that is shown in *Fig. 2-10*. Overplotted this predicted relation are data points from real galaxies.

For those who may have the Sérsic parameter set (R_e, μ_e, n), one can use (2.3) to convert μ_e into $\langle\mu\rangle_e$ if one wishes to compare with the relation given by (2.14). For those who may

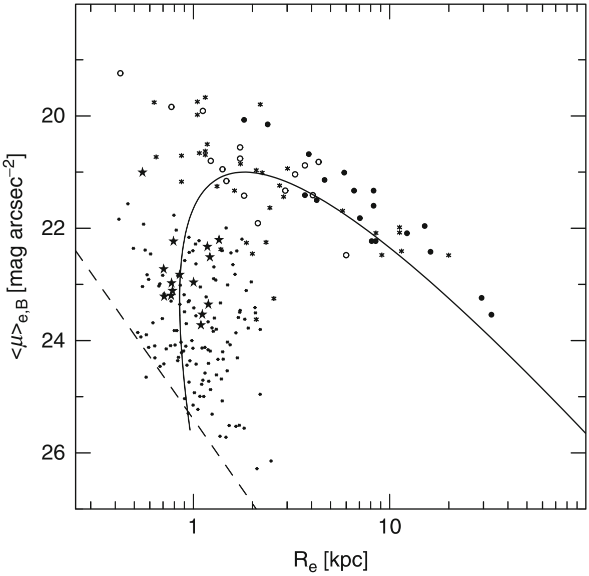


Fig. 2-10

Due to the observed linear relations in *Fig. 2-7*, the relation between the effective radius (R_e) and the mean surface brightness within this radius ($\langle\mu\rangle_e$) is highly curved for elliptical galaxies. The *dashed line* shows the $M_B = -13$ limit for the Virgo cluster

have the parameter set (R_e, μ_e), perhaps obtained with no recourse to the Sérsic model, (2.14) can easily be adjusted using (2.3) to give a relation between R_e and μ_e such that

$$\log R_e = \frac{\mu_e}{15} - 1.81 - 0.5 \log\left(\frac{n e^b \Gamma(2n)}{10^{0.29b}\, b^{2n}}\right), \tag{2.16}$$

where the value of n associated with the value of μ_e is given by

$$\mu_e = 22.8 + 1.086b - 14.1 \log(n). \tag{2.17}$$

To summarize, due to the linear relations in *Fig. 2-7* which connect dwarf and ordinary elliptical galaxies across the alleged divide at $M_B = -18$ (Kormendy 1985b), or at $n = 2$ (Kormendy and Kennicutt 2004), coupled with the smoothly varying change in light profile shape as a function of absolute magnitude, the $\langle\mu\rangle_e$-R_e and μ_e-R_e relations are expected to be curved (*Fig. 2-10*), as previously shown with empirical data by, e.g., Capaccioli and Caon (1991). This also explains why the fitting of a linear relation to (R_e, μ_e) data by Hoessel et al. (1987) resulted in slopes that depended on their galaxy sample magnitude.

The Kormendy relation is a tangent to the bright arm of what is actually a curved distribution defined by the relation given by (2.14) that is taken from Graham (2010). The apparent deviant nature of the dwarf elliptical galaxies from the approximately linear section of the bright end of the $\langle\mu\rangle_e$-R_e distribution does not require that a different physical process be operating. Moreover, as noted by Graham and Guzmán (2004) and Graham (2005), galaxies which appear to branch off at the faint end of the Fundamental Plane (Djorgovski and Davis 1987) – the flat portion at the bright end of a curved hypersurface – also need not have formed from different physical mechanisms. Simulations that assume or reproduce a linear $\langle\mu\rangle_e$-R_e or μ_e-R_e relation, across too great a magnitude range, have failed to mimic the continuous curved distribution defined by real elliptical galaxies. The same is true for simulations of the "Fundamental Plane."

3.3 Broken Relations

3.3.1 Luminosity-(Central Surface Brightness) Relation

While the relation between a galaxy's absolute magnitude and *extrapolated* central surface brightness is remarkably linear (➋ Sect. 3.1.1), there is a clear break in this relation when using the *actual* central surface brightness at the luminous end of the distribution (➋ *Fig. 2-7a*). This departure from the M_B-μ_0 relation by elliptical galaxies brighter than $M_B \approx -20.5$ ($M > 0.5$–$1 \times 10^{11} M_\odot$) was addressed by Graham and Guzmán (2003) in terms of partially depleted cores relative to the outer Sérsic profile (see also Graham 2004; Trujillo et al. 2004; Merritt and Milosavljević 2005; Ferrarese et al. 2006a; Côté et al. 2007). This transition has nothing to do with the alleged divide between dwarf and giant elliptical galaxies at around $M_B = -18$, but is instead directly related with the Sérsic versus core-Sérsic transition at around $M_B = -20.5$.

As noted in ➋ Sect. 2.2.1, such partially depleted cores in luminous core-Sérsic galaxies are thought to have formed from dry, dissipationless galaxy merger events involving the central coalescence of supermassive black holes (Begelman et al. 1980; Ebisuzaki et al. 1991; but see footnote 13) and resulted in Trujillo et al. (2004) advocating a "new elliptical galaxy paradigm" based on the presence of a central stellar deficit versus either none or an excess of light, an approach embraced by Ferrarese et al. (2006a) and Côté et al. (2007), and others.

Further evidence for a division at $M_B = -20.5$ comes from the tendency for the brighter galaxies to be anisotropic, pressure supported elliptical galaxies having boxy isophotes, while the less luminous early-type galaxies may have disky isophotes and often contain a rotating disk (e.g., Carter 1978, 1987; Davies et al. 1983; Bender et al. 1988; Peletier et al. 1990; Jaffe et al. 1994). Core galaxies also tend to be more radio loud and have a greater soft X-ray flux (e.g., Ellis and O'Sullivan 2006; Pellegrini 2010; Richings et al. 2011, and references therein).

It was, in part, a diagram of central surface brightness versus magnitude that led Kormendy (1985b, his Fig. 3) to advocate a separation of dwarf and normal elliptical galaxies at $M_B = -18$ mag. However, as noted by Graham and Guzmán (2003), his sample was missing the bridging population near $M_B = -18 \pm 1$ mag. Excluding galaxies of this magnitude from ➋ *Fig. 2-7a* would also result in two apparently disjoint relations nearly at right angles to each other. Strom and Strom (1978, their Fig. 8; see also Binggeli et al. 1984) had already shown that a linear relation exists between magnitude and central surface brightness from $-18.4 < M_V < -21.6$, spanning the magnitude gap in Kormendy (1985b). Nonetheless, Faber and Lin (1983) had just observed that three of their six dwarf elliptical galaxies had near-exponential light profiles, leading them to speculate that dEs are more closely related to "exponential systems" than (tidally truncated) elliptical galaxies, and Wirth and Gallagher (1984, see also Michard 1979) had also just advocated a division between exponential-dwarf and $R^{1/4}$-giant elliptical galaxies.

To further confound matters, a slope error in Kormendy (1985b) for the distribution of dwarf elliptical galaxies in the M–μ_0 plane meant that, first, the bright end of his dwarf elliptical galaxy distribution did not point toward the faint end of his luminous elliptical galaxy distribution, and thus, there was no suggestion of a connection. Consequences are still seen today (e.g., Tolstoy et al. 2009, their Fig. 1), although Kormendy et al. (2009) have now corrected this. Second, two points representing flattened disk galaxies were added at the bright end of the misaligned dwarf elliptical galaxy sequence by Kormendy (1985b), implying a connection between dwarf elliptical galaxies and disk galaxies rather than ordinary elliptical galaxies.

A decade later, the Astronomy and Astrophysics Review paper by Ferguson and Binggeli (1994; their Fig. 3) had a big question mark as to "how" and indeed "if" dwarf and ordinary elliptical galaxies might connect in this diagram. When galaxies spanning the gap in Kormendy's (1985b) analysis were included by Faber et al. (1997, their Fig. 4c) and shown to follow a trend consistent with the relation from Strom and Strom (1978) and Binggeli et al. (1984), in which the central surface brightness became fainter with decreasing galaxy luminosity, Faber et al. (1997) suggested that this behavior in their data was spurious and due to limited resolution, still using a paradigm in which a discontinuity separated dwarf and ordinary elliptical galaxies. At the same time, Jerjen and Binggeli (1997) argued exactly the opposite, suggesting that it was instead the "core" galaxies which had been displaced in the M–μ_0 diagram from a linear M–μ_0 relation rather than wrong central surface brightness measurements for the faint (non-dwarf) elliptical galaxies. As Graham and Guzmán (2003) and Graham (2004) later explained, in terms of a ~0.1% central mass deficit relative to the outer Sérsic profile in galaxies brighter than $M_B \approx -20.5$, Jerjen and Binggeli (1997) were right, supporting the views expressed by Binggeli et al. (1984) on a continuity between dwarf elliptical and ordinary elliptical galaxies across the alleged divide at $M_B \approx -18$.

An Alternate View

Before continuing, two points regarding the M–μ_0 diagram in Kormendy et al. (2009, their Fig. 34) should perhaps be made. First, Kormendy et al. (2009) have again introduced an apparent gap in their diagram by using only nine galaxies from Binggeli et al.'s (1985) Virgo Cluster Catalog within the magnitude range $M_V = -19 \pm 1$. Obviously, other elliptical galaxies exist with magnitudes over this range, and the HST data in Faber et al. (1997, their Fig. 4c) reveals that these "gap galaxies" – which they identify as "power-law galaxies" – extend the linear $M - \mu_0$ relation defined by the dwarf elliptical galaxies (e.g., Jerjen and Binggeli 1997, their Fig. 3). Faber et al. (1997) did note that these galaxies may have dense, unresolved nuclear components which might increase their central surface brightness and erase this connection. However, the central surface brightness of the Sérsic model describing the host galaxy is not significantly different from that shown in their figure, and as such their speculation, which may be valid, is not actually relevant here as we are interested in the host galaxy unbiased by additional nuclear components.

The second issue regarding Kormendy et al.'s (2009) M–μ_0 diagram is that they have included M32, NGC 4486B plus 4 other "compact elliptical" galaxies fainter than $M_V = -17$ (VCC 1199, VCC 1627, VCC 1440, and VCC 1192; see Binggeli et al. 1985, their Table XIII). They claim that these objects, rather than the non-compact elliptical galaxies without partially depleted cores, define the faint end of the elliptical galaxy sequence. Compact elliptical galaxies were once thought to be small elliptical galaxies because they were, like bright ellipticals, thought to have $R^{1/4}$-like light profiles. However, the bulge-disk nature of M32, the prototype "compact elliptical" which resides close to M31, reveals that it is likely a stripped S0 galaxy (Nieto 1990; Bekki et al. 2001b; Graham 2002) as does the age of its stellar population (Davidge and Beck 2008).[20] Further analysis of other nearby "compact elliptical" galaxies has reaffirmed their bulge/disk nature (e.g., Smith Castelli et al. 2008; Chilingarian et al. 2009; Price et al. 2009;

[20] We speculate that some old, "free-flying" "compact elliptical" galaxies may be the descendants of the compact galaxies seen at $1.5 < z < 2.5$ which simply never accreted and built a significant cold disk (e.g., Kereš et al. 2005).

Huxor et al. 2011). An analysis using CCD data (Graham 2002), rather than the older photographic data used in Kormendy et al. (2009), to define the outer light profile of M32 results in parameters which do not position it at the extension of the core-galaxy distribution in the M–μ_0 diagram.

From the light profile data for VCC 1199 (Kormendy et al. 2009, their Figure 24), a galaxy with the bright companion M49, one can see the bulge/disc transition at 5 arcseconds in the position angle, ellipticity and colour profiles. From the light profile data for VCC 1627 (Kormendy et al. 2009, their Figure 24), one can again see a transition in the position angle and ellipticity profiles, while the colour profile has been truncated. The significant ordered structure in their data minus Sérsic model residual profile reveals the mismatched inner profile shape (evident as the hump out to 13 arcseconds) followed by excess light out to 28 arcseconds. This behaviour is also displayed in the residual profile for VCC 1440, with a transition at about 6 arcseconds, where the ellipticity and position angle also change (Kormendy et al. 2009, their Figure 23). Lastly, there is again clear evidence of this same structure in both VCC 1192's HST/ACS light profile (Kormendy et al. 2009, their Figure 23), another galaxy which resides close to M49, and in NGC 4486B's light profile (Kormendy et al. 2009, their Figure 22), a galaxy which resides close to M87. All of this is suggestive of a bulge/disc nature for these additional "compact elliptical" galaxies.

3.3.2 Luminosity-Color Relation

Additional support for the dissipationless dry-merging scenario at the high-mass end is the flattening of the color-magnitude relation above $0.5\text{–}1 \times 10^{11} M_\odot$. While low luminosity, low Sérsic index, elliptical galaxies are bluer than bright elliptical galaxies (e.g., de Vaucouleurs 1961; Webb 1964; Sandage 1972; Caldwell and Bothun 1987), the brightest galaxies have the same color as each other. This flattening in the color-magnitude relation was noted by Tremonti et al. (2004) and is evident in Baldry et al. (2004, their Fig. 9), Ferrarese et al. (2006a, their Fig. 123), Boselli et al. (2008, their Fig. 7), and even Metcalfe et al. (1994). These observations help alleviate past tension with semi-analytic models that had predicted a relatively flat color-magnitude relation for bright elliptical galaxies (e.g., Cole et al. 2000). Previously, based on what was thought to be a linear color-magnitude relation, Bernardi et al. (2007) had written that "if BCGs formed from dry mergers, then BCG progenitors must have been red for their magnitudes, suggesting that they hosted older stellar populations than is typical for their luminosities." However, the flattening in the color-magnitude relation has since been recognized in yet more data sets (e.g., Skelton et al. 2009; Jiménez et al. 2011), although it should perhaps be noted that Skelton et al. reported the transition at $M_R = -21$, i.e., ~1 mag fainter.

In passing, it is noted that the relation between luminosity and supermassive black hole mass (Marconi and Hunt 2003; McLure and Dunlop 2004) as refined by Graham (2007) shows a linear relation for black hole masses predominantly greater than $10^8 M_\odot$ – consistent with the concept of dry galaxy merging at this high-mass end.

3.3.3 Dynamics

From a sample of 13 early-type galaxies, plus one spiral galaxy, Minkowski (1962) noted that a "correlation between velocity dispersion and (luminosity) exists, but it is poor." He wrote that "it seems important to extend the observations to more objects, especially at low and medium absolute magnitudes." This was done by Morton and Chevalier (1973) who noted the same "continuous distribution of dispersions from 60 km/s for M32 to 490 km/s for M87" but also did not attempt to quantify this trend. It was Faber and Jackson (1976) who, with improved data and a larger sample of 25 galaxies, were the first to quantify Minkowski's relation and discovered that $L \propto \sigma^4$ for their data set. This result has proved extremely popular and is known as the Faber-Jackson relation. Not long after this, Schechter (1980) and Malumuth and Kirshner (1981) revealed that the luminous elliptical galaxies followed a relation with an exponent of ~5 rather than 4. At the same time, Tonry (1981) revealed that expanding the sample to include more faint elliptical galaxies results in an exponent of ~3. This led Binney (1982) to write that "probably the correlation cannot be adequately fitted by a single power law over the full range of absolute magnitudes," and Farouki et al. (1983) wrote that "the data suggests the presence of curvature in the $L - \sigma$ relation." Davies et al. (1983), and later Held et al. (1992), showed that the dwarf elliptical galaxies followed a relation with an exponent of ~2, which explains why Tonry (1981) had found a slope of ~3 when including dwarf and ordinary elliptical galaxies. The relation found by Davies et al., with a slope of ~2, has recently been observed by de Rijcke et al. (2005), and the curved or possibly broken $L - \sigma$ distribution has been interpreted by Matković and Guzmán (2005) as a change in slope at $M_B = -20.5$ (see also Evstigneeva et al. 2007) in agreement with Davies et al. (1983).

In spite of all the above work, there is a huge body of literature today which does not reflect the evidence for an $L - \sigma$ relation that is curved or broken. Simulations of galaxies which succeed in producing the linear Faber-Jackson relation, $L \propto \sigma^4$, have actually failed to produce the full distribution of dynamics seen in real elliptical galaxies as a function of magnitude.

4 Disk Galaxy Light Profiles

The term "disk galaxy" refers to galaxies with large-scale stellar disks, encompassing both spiral (Sp) and lenticular (S0) galaxies, with the latter not displaying a noticeable spiral pattern of stars nor ongoing star formation – at least not in the optical (Gil de Paz et al. 2007). Broadly, their bulges are centrally located stellar distributions with a smooth appearance, which appears as an excess, *a bulge*, relative to the inward extrapolation of the outer exponential disk light (excluding bars, nuclear star clusters, and nuclear disks). Readers may like to refer to Wyse et al. (1997) for a fuller definition.

It is perhaps worth noting that as galaxy photometry and image analysis have improved over the years, along with the availability of kinematic information, it has become increasingly apparent that some galaxies previously labeled as elliptical galaxies actually possess a large-scale stellar disk and are in fact lenticular galaxies (Davies et al. 1983; Bender et al. 1988; Rix and White 1990). These have long been known to occur among the fainter early-type galaxies, rather than the massive elliptical galaxies (e.g., Liller 1960, 1966). Similarly, albeit to a lesser extent, some previously classified dwarf elliptical (dE) galaxies have also been recognized to contain a large-scale stellar disk and are therefore actually dwarf lenticular (dS0), and rarely dwarf spiral (dSp) galaxies (e.g., Graham and Guzmán 2003; Lisker et al. 2006a). To accurately quantify the optical structure of these, and all, disk galaxies, require, as a minimum, that one models the 3D bulge (when present) and the flattened 2D disk as distinct entities rather than fitting a single $R^{1/n}$, or worse still $R^{1/4}$, model to both components.

The following section reviews the decomposition of disk galaxy light into its two primary (bulge and disk) components. A brief discussion regarding the problems of dust is provided in Sect. 4.2, while a discussion of the problems with the identification of pseudobulges is presented in Sect. 4.3. A few words about bulgeless galaxies are provided in Sect. 4.4 while a number of references to works which have advanced the modeling, and shown the importance, of bars are provided in Sect. 4.5. Those interested in the more detailed morphology and components of disk galaxies may like to look at Chap. 1 in this volume.

4.1 The Bulge-Disk Decomposition

It has long been known that the (azimuthally averaged) radial distribution of starlight in disks (i.e., the disk component of disk galaxies) can be reasonably well approximated with an exponential model (e.g., Patterson 1940; de Vaucouleurs 1957, 1958; Freeman 1970), such that the intensity I varies with R according to the expression

$$I(R) = I_0\, e^{-R/h}, \tag{2.18}$$

where I_0 is the central intensity and h is the e-folding disk scale length. Expression relating these two quantities to the parameters of the Sérsic $R^{1/n}$ model, which reproduces an exponential model when $n = 1$, can be found in Graham and Driver (2005).

While the intensity profiles of these disks decline with radius, uninterrupted in most galaxies (e.g., Barton and Thompson 1997; Weiner et al. 2001; Bland-Hawthorn et al. 2005), deviations from this general exponential disk structure have also been known for a long time. Spiral arms can introduce bumps and upturns in the light profile, as seen in the data from de Jong and van der Kruit (1994) and highlighted by Erwin et al. (2005). On the other hand, some disks are known to partially truncate at a few disk scale lengths (van der Kruit 1987; Pohlen et al. 2004, and references therein), including our own (Minniti et al. 2011), often resulting in an apparent double-disk structure (van der Kruit 1979; van der Kruit and Searle 1981). Due to the nonzero thickness of disks, one can additionally study their vertical structure orthogonal to the plane of the disk (van der Kruit and Searle 1981; van der Kruit 1988; de Grijs et al. 1997; Qu et al. 2011, and references therein). Although it has been recognized that some galaxies may have both a thin and a thick disk (Gilmore and Reid 1983; Chang et al. 2011), the level of observational detail in most galaxies makes this separation impossible, and as such, it is not common practice to attempt this differentiation in external galaxies.

Around the time of Capaccioli's (1985, 1989) advocation of Sérsic's $R^{1/n}$ model, Shaw and Gilmore (1989) and Wainscoat et al. (1989) remarked that the bulge component of spiral galaxies is neither all similar nor are they adequately described by the $R^{1/4}$ model. This was not a new result, as de Vaucouleurs (1959) had himself noted departures from the $R^{1/4}$ model, and van Houten (1961) had clearly demonstrated that an exponential model provided a better fit to some bulges than the $R^{1/4}$ model (see also, e.g., Liller 1966, his Fig. 2; Frankston and Schild 1976; Spinrad et al. 1978). The exponential model was also eventually shown to provide a good description for the bulge of the Milky Way (Kent et al. 1991). While Andredakis and Sanders (1994) demonstrated that many bulges are better fit with an exponential model than an $R^{1/4}$ model, it was Andredakis et al. (1995) who demonstrated that Sérsic's $R^{1/n}$ model provided a good description for the bulges of all disk galaxies. They discovered that the concentration of stars in bulges, quantified by the Sérsic index n, varied with bulge mass. After a number of other early works by Heraudeau and Simien (1997), Iodice et al. (1997, 1999), Schwarzkopf and

Table 2-1

***K*-band structural parameters for spiral galaxies, as a function of morphological type (Graham and Worley 2008). The median bulge Sérsic index is listed, along with the median bulge-to-disk size ratio R_e/h, and the median dust-corrected bulge-to-disk (B/D) luminosity ratio. The bulge-to-total (B/T) luminosity ratio can be obtained from the expression $B/T = [1 + (D/B)]^{-1}$. The range shown represents ±34% of the distribution about the median**

Type	Sérsic n	R_e/h	$\log(B/D)$
Sa	$2.56^{+2.79}_{-0.79}$	$0.31^{+0.20}_{-0.17}$	$-0.34^{+0.40}_{-0.32}$
Sab	$2.45^{+1.27}_{-0.75}$	$0.24^{+0.22}_{-0.10}$	$-0.54^{+0.53}_{-0.41}$
Sb	$2.00^{+1.62}_{-0.76}$	$0.21^{+0.15}_{-0.07}$	$-0.60^{+0.28}_{-0.49}$
Sbc	$1.87^{+1.64}_{-0.75}$	$0.21^{+0.11}_{-0.09}$	$-0.82^{+0.28}_{-0.42}$
Sc	$1.78^{+2.18}_{-0.79}$	$0.22^{+0.27}_{-0.09}$	$-1.06^{+0.43}_{-0.34}$
Scd	$1.18^{+0.89}_{-0.49}$	$0.19^{+0.10}_{-0.06}$	$-1.23^{+0.75}_{-0.28}$
Sd	$1.80^{+0.49}_{-1.22}$	$0.24^{+0.07}_{-0.10}$	$-1.06^{+0.16}_{-0.50}$
Sdm	$0.79^{+0.24}_{-0.13}$	$0.19^{+0.14}_{-0.07}$	$-1.49^{+0.36}_{-0.36}$
Sm	$0.40^{+0.09}_{-0.09}$	$0.23^{+0.01}_{-0.01}$	$-1.57^{+0.01}_{-0.01}$

Dettmar (1997), Seigar and James (1998), and Wadadekar et al. (1999), it has become common practice to routinely fit disk galaxies with a Sérsic bulge plus an exponential disk. Typical bulge Sérsic indices and bulge-to-disk flux ratios for spiral galaxies are provided in *Table 2-1*.

This Sérsic-bulge plus an exponential-disk approach was used by Allen et al. (2006) to model over 10,000 galaxies from the Millennium Galaxy Catalogue (Liske et al. 2003). From this, the luminosity function of bulges and disks, rather than simply galaxies, were constructed (Driver et al. 2007a). This is important because the spheroidal bulge and flattened-disk component of galaxies formed through different physical processes. If one is to more fully understand the growth of galaxies, tracing their evolution in simulations or observing their evolution over a range of redshifts, then the disk galaxies should not be treated as single-component systems. The stellar mass density of bulges and disks was subsequently derived by Driver et al. (2007a), superseding studies with (a) notably smaller sample sizes and (b) that had used $R^{1/4}$ models. A full correction for the obscuring effects of dust was provided by Driver et al. 2007b).

4.2 Dust and Inclination Corrections

The impact of interstellar dust[21] (see the review by Draine 2003) often does not receive the attention that it deserves. In addition to Trimble's (1999) scholarly report on how most studies from circa 1850 until Trumpler (1930a, b) ignored reports that the space between stars is not transparent, the influence of dust on modern, UV-optical, and near-infrared measurements of the luminosity of bulges and disks is still to receive due diligence. This is not due to a complete lack of appreciation but rather because it is a difficult topic which has until recently been swept under the rug.

Although the importance and extent of dust's thermal *emission* at infrared wavelengths have been studied in external galaxies for over two decades (e.g., Devereux and Young 1990; Goudfrooij and de Jong 1995; Temi et al. 2004), the importance of dust *obscuration* at UV, optical, and

[21] Dust was formerly referred to as "dark matter" (e.g., Trumpler 1930a, b).

near-IR wavelengths has not fully filtered through to the entire community – even though it is primarily the re-radiation of this absorbed energy which produces the infrared and millimeter wavelength emissions. It is hoped that the level of dust recognition may soon increase due to the vast interest and awareness in results from Spitzer (Werner et al. 2004) and coming from the infrared Herschel Space Observatory (Pilbratt et al. 2010; see also Walmsley et al. 2010) and also from the sub-mm Planck mission (e.g., Ade et al. 2011).

While most elliptical galaxies do not possess much diffuse dust (Bregman et al. 1998; Clemens et al. 2010), lenticular and spiral disk galaxies can have copious amounts (e.g., White et al. 2000; Keel and White 2001a, b). As a result, the far side of a bulge viewed through the dusty disk component of a disk galaxy can have its light substantially diminished. Corrections for dimming due to dust, beyond the necessary masking in one's image when faced with obvious dust lanes and patches (e.g., Hawarden et al. 1981), therefore need to be applied to both the disk and bulge components of disk galaxies. To date, at optical wavelengths, these corrections have typically been inadequate for the disk and nonexistent for the bulge (e.g., de Jong 1996; Graham 2001; Unterborn and Ryden 2008; Maller et al. 2009).

A simple schematic often used to help explain the influence of dust on disks is to start by considering an optically thick, face-on disk, in which you can only see some depth l into the disk before your view is obscured by dust. As this disk is inclined by some angle i, toward an edge-on orientation $i = 90^\circ$, your line-of-sight depth into (the z direction of) the disk will be reduced by $\cos(i)$, and thus you will see even less stars, with the observed disk luminosity declining as the inclination increases toward an edge-on orientation. Numerous studies have revealed how the observed luminosity of disk galaxies declines with inclination (e.g., Tully et al. 1998; Masters et al. 2003). Correcting for this "inclination effect," in a sample average sense, yields the galaxy luminosities that would be observed if the inclined galaxies that one observed had a face-on orientation, and this is typically the only correction applied. However, after correcting to a face-on orientation, i.e., after determining the amount of flux that one would observe if the disk galaxies were seen face-on, one still needs to apply an additional correction to determine the luminosity that would be observed from the face-on disk galaxies if the dust was not present. To date, most observational studies have only calibrated disk galaxy magnitudes to the face-on orientation; they have not then accounted for the obscuring dust which remains. To do so requires recourse to sophisticated radiative transfer models.

One such model is that from Popescu et al. (2000, 2011) which self-consistently explains the UV/optical/FIR/sub-mm emission from galaxies. Other recent radiative transfer codes which readers may find helpful include TRADING (Bianchi 2008), GRASIL (Schurer et al. 2009), and SUNRISE (Jonsson 2006; Jonsson et al. 2010). Using these models, a renaissance of sorts is slowly emerging in regard to correcting the optical, and near-infrared, luminosities of disk galaxies. Expressions are now available to fully correct for dust extinction in disk galaxies (e.g., Driver et al. 2008; Graham and Worley 2008). Such equations will become more refined as dust corrections for different disk morphological types, or luminosity, eventually become available.

Properly accounting for dust at ultraviolet, optical, and near-infrared wavelengths will have many ramifications. For example, the dust-corrected ultraviolet flux density in the local Universe (Robotham and Driver 2011) is significantly greater than previously reported, as is the associated star formation rate (SFR) density which is now $0.0312 \pm 0.0045\, h\, M_\odot\, \text{year}^{-1}\, \text{Mpc}^{-3}$. We are also now in a position to construct dust-free Tully and Fisher (1977) relations, rather than relations that are based on the luminosities of dusty galaxies corrected only to a face-on inclination. Observations of the redshift evolution of this dust correction remain unexplored. Until addressed, measurements of evolution in galaxy luminosities and luminosity functions remain somewhat limited (although see the simulations by Somerville et al. (2011)). Finally, it is also noted that one should correct for inclination and face-on dust extinction, if only implicitly through template spectra, and if using (the increasingly popular) photometric measurements to estimate redshifts (e.g., Yip et al. 2011).

Correcting for dust in galaxies has revealed that the luminosity output of the (local) Universe, at visible wavelengths from starlight, is actually twice as bright as observed (Driver et al. 2008; see also Soifer and Neugebauer 1991). It has also enabled a more accurate estimate of the stellar mass density of bulges and disks in the Universe today (i.e., at $z = 0$), with Driver et al. (2007b) reporting that $\rho_{\text{disks}} = 4.4 \pm 0.6 \times 10^8\, hM_\odot\, \text{Mpc}^{-3}$ and $\rho_{\text{bulges}} = 2.2 \pm 0.4 \times 10^8\, hM_\odot\, \text{Mpc}^{-3}$, implying that $11.9 \pm 1.7\, h$ percent of the baryons in the universe (Salpeter-"lite" IMF) are in the form of stars, with ≈58% in disks, ≈10% in red elliptical galaxies, ≈29% in classical bulges, and the remainder in low luminosity blue spheroidal systems.

4.3 Pseudobulges

Pseudobulges are rather controversial, and the final section of Wyse et al. (1997) highlights many concerns. These exponential-like bulges, formed from disk material, were discussed at length in Hohl (1975, see his Fig. 6) which covered results from N-body simulations of disks over the previous 7 years (see also Bardeen 1975). While it has since been shown that such "pseudobulges" can coexist with "classical" bulges in the same galaxy (e.g., Norman et al. 1996; Erwin et al. 2003; Athanassoula 2005), pseudobulges remain difficult to identify (because, in part, we now know that classical bulges can also have exponential light profiles), and they bring to mind different classes of object for different authors. Even in our own galaxy, there has been confusion as to whether a pseudobulge or a classical bulge exists (Babusiaux et al. 2010). Due to an increasing number of papers using questionable criteria to separate bulges into either a classical bulge or a pseudobulge bin, this section has been included. However, readers who may regard the identification of pseudobulges as pseudoscience at this time may wish to skip to the following subsection.

Pseudobulges are not thought to have formed from violent relaxation processes such as monolithic collapse, early rapid hierarchical merging, nor late-time merger events nor are they thought to have formed from cluster harassment of disk and irregular galaxies, i.e., processes that may have built the elliptical galaxy sequence. Such processes give rise to what are termed "classical bulges." Pseudobulges – or disky bulges in the notation of Athanassoula (2005) – are comprised of, or built from, disk material through either secular evolutionary processes (Hohl 1975; Combes and Sanders 1981; Combes et al. 1990) or alternatively invoked by external triggers (Mihos and Hernquist 1994; Kannappan et al. 2004). Pseudobulges are supposed to rotate and have an exponential light profile, akin to the disk material from which they formed (Hohl and Zhang 1979; Kormendy 1982; Pfenniger 1993; Kormendy and Kennicutt 2004), although, as we shall see, such behavior does not confirm their existence.

For many years, the bulges of disk galaxies were generally thought to resemble small elliptical galaxies displaying an $R^{1/4}$ light-profile (e.g., de Vaucouleurs 1958; Kormendy 1977a; Kent 1985; Kodaira et al. 1986; Simien and de Vaucouleurs 1986). After Andredakis and Sanders' (1994) revelation that many bulges are better approximated with an exponential model, some studies, not ready to abandon the $R^{1/4}$ model, started to report on an apparent bulge dichotomy: classical $\text{R}^{1/4}$ bulges versus exponential bulges allegedly built from the secular evolution of the disk. After the eventual recognition that a continuous range of bulge profile shapes and properties exist (e.g., Andredakis et al. 1995; Khosroshahi et al. 2000; Graham 2001; Möllenhoff and Heidt 2001; MacArthur et al. 2003), we have since seen the practice by some of dividing classical versus secular bulges upon whether their Sérsic index n has a value greater than or less than 2 (Kormendy and Kennicutt 2004; Fisher and Drory 2008).

Kormendy and Kennicutt (2004) advocate several additional criteria, which sometimes can be used and other times cannot, for identifying pseudobulges. Their first criteria is a flattened shape similar to a disk, which is arguably the best criterion in theory but often difficult to observe in practice. Kormendy and Kennicutt claim that the presence of an inner spiral pattern (or a nuclear bar) rules out the presence of a classical bulge. However, spiral patterns (in disks) can apparently exist within classical bulges (Jerjen et al. 2000b; Thakur et al. 2009). Moreover, the existence of inner disk within elliptical galaxies (e.g., Rest et al. 2001; Tran et al. 2001) is evidence that a classical bulge may be present, surrounding the disk. Indeed, Kormendy et al. (2005) reiterate past suggestions that secular growth within a classical bulge may explain such inner disks, a scenario further supported by Peletier et al. (2007). That is, rather than secular evolution within a disk building a pseudobulge, one may have the construction of a disk within a classical bulge. A third criterion from Kormendy and Kennicutt (2004) is that a boxy bulge in an edge-on galaxy rules out a classical bulge. Williams et al. (2010) have however shown that boxy bulges (previously), thought to be bars seen in projection (Combes and Sanders 1981), do not all display cylindrical rotation and that they can have stellar populations different from their disk, revealing the presence of a classical (and very boxy) bulge. Furthermore, Binney and Petrou (1985) established that mergers can build boxy, cylindrically rotating bulges. As noted, Sérsic index and rotation, reflecting the results from Hohl (1975), are additionally offered as a tool by Kormendy and Kennicutt (2004) to separate pseudobulges from classical bulges, and these criteria are addressed separately below, along with a discussion of ages and structural scaling relations. The intention is not to argue that "pseudobulges" do not exist but only to raise awareness that there are substantial grounds to question their identification from certain selection criteria.

4.3.1 Sérsic Index

It has been suggested that one can identify pseudobulges if they have a Sérsic index of 1, i.e., an exponential light profile, or, more broadly, if they have a Sérsic index less than 2 (Kormendy and Kennicutt 2004). Given that bulges with $R^{1/4}$-like profiles are now known to be relatively rare (Andredakis and Sanders 1994) even among early-type disk galaxies (Balcells et al. 2003), this approach is somewhat problematic as it would tend to identify the majority of bulges as pseudobulges based upon the 400+ bulge Sérsic indices tabulated by Graham and Worley (2008). Given the extent of galaxy merging predicted by hierarchical models, it seems unlikely that most disk galaxies have not experienced a merger event that has influenced their bulge. We argue that use of the Sérsic index to identify pseudobulges is at best risky and at worst highly inappropriate.

As we have seen in this article, elliptical galaxies display a continuum in numerous properties as a function of mass (see also Graham and Guzmán 2003; Côté et al. 2007, and references therein). One of these properties is the concentration of the radial distribution of stars, i.e., the shape of the light profile (e.g., Caon et al. 1993; Young and Currie 1994). The low-mass dwarf elliptical galaxies, not believed to have formed from the secular evolution, or perturbation, of a disk, have Sérsic indices around 1–2. Therefore, disk galaxy bulges with Sérsic indices $n < 2$ need not have formed via secular disk processes, although some may have. The bulge of the Milky Way is a good example where an exponential (Sérsic $n = 1$) model (Kent et al. 1991; Graham and Driver 2007) describes the light of what is a classical bulge (Babusiaux et al. 2010). Furthermore, while secular evolution is not proposed to build large bulges, like the one in the Sombrero galaxy, such big bulges display a continuum of profile shapes with ever smaller bulges, such that the Sérsic index decreases as the bulge luminosity decreases (➋ *Table 2-1*), which has led many to speculate that a single unifying physical process is operating (e.g., Peletier and Balcells 1996b) or at least overriding what has gone before.

Domínguez-Tenreiro et al. (1998) and Aguerri et al. (2001) have grown bulges from hierarchical simulations and minor merger events that have Sérsic indices from 1 to 2. Scannapieco et al. (2011) have also shown that classical bulges formed by mergers can have Sérsic indices less than 2. Although their gravitational softening length was only two to four times their bulge scale lengths, this is comparable to the situation that observers often deal with (e.g., de Jong 1996). By fitting simulated light profiles, Gadotti (2008) revealed that one can reliably measure the Sérsic index for bulges if their half-light radius is larger than just 80% of the image's point-spread function's (PSF's) half width half maximum (HWHM).

4.3.2 Rotation

The identification of rotating bulges goes back a long time (e.g., Pease 1918; Babcock 1938, 1939; Rubin et al. 1973; Pellet 1976; Bertola and Capaccioli 1977; Peterson 1978; Mebold et al. 1979). Early-type galaxies can also display significant rotation, albeit likely due to the presence of a disk (e.g., Davies et al. 1983; Graham et al. 1998; Emsellem et al. 2007, 2011; Krajnović et al. 2008). Due to the presence of a bar, as opposed to a pseudobulge, classical bulges can also appear to rotate (e.g., Babusiaux et al. 2010). Classical bulges can also be spun up by a bar (Saha et al. 2011). To further complicate matters, simulations indicate that merger events can result in rotating elliptical galaxies (Naab et al. 1999, 2006; González-García et al. 2009; Hoffman et al. 2009), and Bekki (2010) has shown that a rotating bulge can also be created by a merger event and is thus not necessarily a sign of a pseudobulge built by secular evolution from disk material as typically assumed. In addition, Qu et al. (2011) report on how the rotational delay between old and young stars in the disc of our Galaxy may be a signature of a minor merger event.

This is not to say that inner, rotationally flattened, and supported disks do not masquerade as bulges (e.g., Erwin et al. 2003), only that rotation may not be a definitive sign of "bulges" built via secular disk processes.

4.3.3 Ages

The bulge of the Milky Way consists of an old stellar population (Rich and Origlia 2005; Zoccali et al. 2006; Cavichia et al. 2012) and has kinematics consistent with a pressure-supported system rather than a rotationally supported system (Wyse and Gilmore 1995; Babusiaux et al. 2010). Contrary to classical bulges formed early on in the universe, pseudobulges are thought to be young, built from disk material.

Using optical and near-infrared colors, Bothun and Gregg (1990) had argued that S0 galaxy disks were more than 5-Gyr younger than the bulges they host, affirming the previously held belief that bulges are akin to old elliptical galaxies (Renzini 1999). With refined measurements, and avoiding obvious dusty regions, Peletier and Balcells (1996a) discovered that the color and age difference (assuming old populations with identical metallicities) between the bulge and disk from the same galaxy, in a sample of early-type disk galaxies, are much closer than had been realized, but still with the bulges older than their surrounding disks. Peletier et al. (1999) went on to conclude that the bulges of their S0-Sb galaxies are indeed old and cannot have formed by secular evolution[22] more recently than $z = 3$ (see also Goudfrooij et al. 1999, and Bell and de Jong 2000). From a sample of nine late-type galaxies, Carollo et al. (2007) used optical and near-infrared images to discover that roughly half of their bulges were old and half were young (see also Gadotti and Anjos 2001).

Dust and bright young blue stars ($\lesssim$1 Gyr) can significantly bias the light at optical wavelengths (MacArthur et al. 2010), as can young-ish ($\lesssim$2 Gyr) stars in the near-infrared due to thermally pulsing asymptotic giant branch stars (e.g., Freeman 2004; Tonini et al. 2010). From a line strength analysis, Thomas and Davies (2006) concluded that secular evolution is not a dominant mechanism for Sbc and earlier-type spirals, and Moorthy and Holtzman (2006) concluded that merging rather than secular evolution is likely the dominant mechanism for bulge formation (see also Jablonka et al. 2007). MacArthur et al. (2009) revealed, also with spectra rather than colors, that bulges in both early- and late-type spiral galaxies, even those with Sérsic indices $n < 2$, have old mass-weighted ages, with less than 25% by mass of the stars being young. Based on the bulge's stellar populations and stellar gradients (see also Fisher et al. 1996), they concluded that early-formation processes are common to all bulges and that secular processes or "rejuvenated" star formation generally contributes minimally to the stellar mass budget but has biased luminosity-weighted age estimates in the past. Such "frostings" of young stars, up to some 25%, can bias some techniques into missing the dominance of old stars by mass in (most) bulges.

4.3.4 Scaling Relations

Following on from *Fig. 2-7* for elliptical galaxies, *Fig. 2-11* reveals how the K-band magnitudes of nearly 400 bulges vary with (a) Sérsic index and (b) the central surface brightness (extrapolated from the Sérsic fit outside of the core). The data points have come from the compilation by Graham and Worley (2008). While the scatter is large, there is no evidence of a discontinuity between bulges with n greater than or less than 2. The following linear relations, shown in the figures, appear to roughly describe the distributions:

$$M_K = -7.5 \log(n) - 20.0; \tag{2.19}$$

$$M_K = 0.6\mu_{0,K} - 29.7. \tag{2.20}$$

The curved relation shown in *Fig. 2-11c* is not a fit but simply

$$M_K = 0.6(\mu_{e,K} - 1.086b) - 29.7, \tag{2.21}$$

where $b \approx 1.9992n - 0.3271$ and $n = 10^{-(20.0+M_K)/7.5}$. That is, the above two linear relations predict the existence of this curved magnitude (effective surface brightness) relation.

The scatter in the data, and the lack of bulges with $n > 4$, makes the curvature in *Fig. 2-11c* hard to see with the current data. Instead, it may *appear* that there is a cascade of fainter data

[22] Peletier et al. (1999) additionally observed three late-type spirals with blue colors indicative of a young luminosity-weighted age.

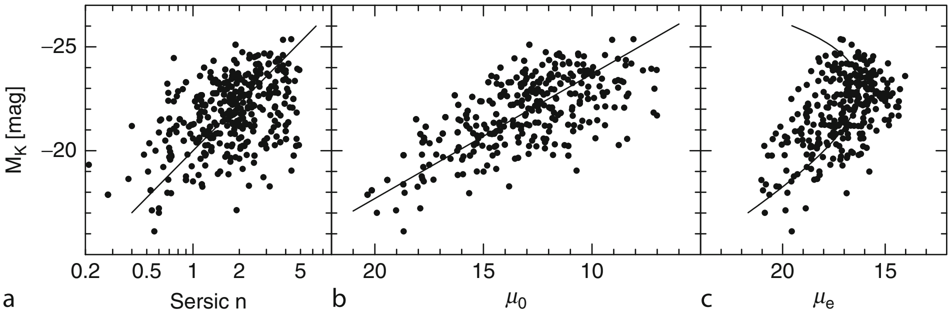


Fig. 2-11

***K*-band bulge magnitude versus (a) Sérsic index, (b) central surface brightness (from the best-fitting Sérsic model), and (c) effective surface brightness (compare *Fig. 2-8b*) (The data have come from the disk-galaxy compilation by Graham and Worley (2008), and parameters were available for most bulges in each panel)**

departing from the region associated with the bright arm of the predicted relation corresponding to bulges with larger Sérsic indices. This apparent departure of fainter galaxies does not, however, imply that two physical processes must be operating.

Due to bulges possessing a continuous range of stellar concentrations, with their luminosity-dependent Sérsic indices n ranging from less than 1 to greater than 4 (*Fig. 2-11a*), several bulge parameters will follow a number of nonlinear scaling relations. The explanation for how this arises is provided in greater detail in Sect. 3.2. Departures from the magnitude-(effective surface brightness) relation, the Kormendy (1977b) relation, and other relations including the Fundamental Plane (Djorgovski and Davis 1987) are expected at the low-luminosity, low Sérsic index, faint central surface brightnesses region of the bulge sequence. This does not imply that these bulges are pseudobulges (Greene et al. 2008; Gadotti 2009), although some may be.

4.4 Bulgeless Galaxies

The bulge-to-disk flux ratio, known to vary with disk morphological type (see *Table 2-1*) is primarily due to changes in the bulge luminosity along the spiral galaxy sequence, as first noted by Yoshizawa and Wakamatsu (1975, their Figs. 1 and 2) and reiterated by Ostriker (1977). Are there disk galaxies at the (small-bulge)-end of the sequence which actually have no bulge?

Although a central "bulge" in the light profile of the Triangulum nebula, M33, has long been evident (e.g., Stebbins and Whitford 1934, and references therein; van den Bergh 1991; Wyse et al. 1997), it has become common practice by many to nowadays refer to M33 as bulgeless. This is a reflection of efforts by some to try and distinguish some excesses of central light relative to the inward extrapolation of the outer exponential disk from a "classical" bulge of stars (e.g., Böker et al. 2003; Walcher et al. 2005). However, some papers have even started referring to the Milky Way as bulgeless, because they feel that it may have a pseudobulge built from disk instabilities and also labeling any galaxy whose bulge's Sérsic index is less than 2 as "pure disk" galaxies. As stressed by Cameron et al. (2009, their Sect. 4), and Allen et al. (2006) – who modeled over 10,000 galaxies with a Sérsic bulge plus an exponential disk – may have added to the confusion by referring to galaxies in which the bulge could not be resolved, and was thus simply ignored in the fit, as "pure disk" galaxies (see also Kautsch 2009a) rather than labeling them "quasi-bulgeless" as done by Barazza et al. (2008). Obviously, the above practices have lead to quite a confusing situation in the literature today. Nonetheless, there are some good examples of what most would agree are almost truly bulgeless galaxies (free from either a classical or a pseudobulge), such as IC 5249 and NGC 300 with less than 2% bulge light (Bland-Hawthorn et al. 2005) and the superthin edge-on galaxies reviewed by Kautsch (2009b).

Current interest in bulgeless galaxies exists for at least two reasons. The first reason extends to disk galaxies with small bulges. Most galaxy simulations have, in the past, had a tendency to produce bulges rather than (pure) disks because of baryon angular momentum losses during major merger events (e.g., Navarro and Benz 1991; D'Onghia et al. 2006) and due to central star formation from minor merger events (Stewart et al. 2008). As highlighted by Graham and Worley (2008) and Weinzirl et al. (2009), many disk galaxies have small (<1/4) bulge-to-total flux ratios (see ➋ *Table 2-1*), which has been at odds with most simulations until recently (Koda et al. 2009; Governato et al. 2010; Brook et al. 2011; Fontanot et al. 2011).

The second reason pertains to claims of supermassive black holes (SMBHs) in bulgeless galaxies. Given the apparent wealth of data revealing correlations between the masses of supermassive black holes and the properties of their host bulge (e.g., Ferrarese and Ford (2005)), the existence of supermassive black holes in bulgeless galaxies is an interesting unresolved problem. While the central kinematic data for the "bulgeless" galaxy M33 is consistent with no SMBH (Merritt et al. 2001; Gebhardt et al. 2001), SMBHs in allegedly bulgeless galaxies includes NGC 4395 (Filippenko and Ho 2003, but see Graham 2007), NGC 1042 (Shields et al. 2008, but see Knapen et al. 2003), NGC 3621 (Gliozzi et al. 2009, but see Barth et al. 2009, their Fig. 4), NGC 3367, and NGC 4536 (McAlpine et al. 2011, but read their text and see Dong and De Robertis 2006). However, if the bulges of these "bulgeless" galaxies are pseudobulges, then questions arise as to the origin and connection of the SMBHs with these bulges. Given the connection between pseudobulges and bars (Hohl 1975), it is pertinent to perform bulge-bar-disk decompositions in galaxies with prominent bars, where the bar is perhaps still more connected to the disk instability than the ensuing pseudobulge (Graham 2011).

4.5 Barred Galaxies

Disk galaxies can possess more components than just a bulge and disk; they may, for example, have bars, oval distortions and lenses, inner and outer rings, spiral arms, local star formation, etc. Features arising from global instabilities in the disk are reviewed in ➋ Chap. 18 of Volume 5. As was noted by the Third Reference Catalogue of de Vaucouleurs et al. (1991, and references therein), multiple components often exist, and sometimes these have been modeled (e.g., Tsikoudi 1979; Prieto et al. 1997, 2001). In addition to these large-scale features, nuclear bars and disks can also be present (e.g., de Zeeuw and Franx 1991; Rest et al. 2001), and when the resolution permits it, these can also be modeled (e.g., Balcells et al. 2007; Seth et al. 2008).

In this section, we shall, however, only briefly discuss large-scale bars (reviewed by Knapen 2010, see also Buta's article in this volume). Roughly half to three-quarters of disk galaxies display a large-scale bar (e.g., de Vaucouleurs 1963; Eskridge et al. 2000; Marinova et al. 2012, and references therein), and many papers have performed a quantitative analysis of these (e.g., Martin 1995; de Jong 1996; Aguerri et al. 1998). Accounting for bars is important where these features are of sufficiently large amplitude to influence the models being fit to the light distribution (e.g., Laurikainen et al. 2005, 2007; Reese et al. 2007; Gadotti 2008; Weinzirl et al. 2009).

Ferrers' (1877) ellipsoid is often used to describe bars (e.g., Sellwood and Wilkinson 1993), as is Freeman's (1966) elliptical cylinder (e.g., de Jong 1996), generalized by Athanassoula et al. (1990). Sérsic's model is also sometimes used with $n = 0.5$, which reproduces a Gaussian function and has a sharpish decline at large radii for approximating the behavior seen at the ends of bars, although truncated models are also sometimes employed. While the bar shows up as a plateau in the galaxy light-profile corresponding to the bar's major-axis, in some galaxies, the azimuthally averaged light profile effectively recombines the bar and (apparently hollowed inner) disk light to reproduce an exponential profile, like that seen in the disks of non-barred galaxies (e.g., Ohta et al. 1990; Elmegreen et al. 1996).

The strength of a bar is often quantified by its ellipticity (e.g., Athanassoula 1992; Martin 1995; Wozniak et al. 1995) – a measure of the departure from the circular orbits of the disk stars – rather than its luminosity contrast (e.g., Ohta et al. 1990; Rozas et al. 1998). Strong bars have major-to-minor axis ratios of 3–4 while very strong bars may have ratios as high as 5. The gravitational torque has also been used to quantify bar strength (e.g., Combes and Sanders 1981; Laurikainen and Salo 2002) and similarly in triaxial bulges (e.g., Trujillo et al. 2002).

5 Summary

We have reviewed the progress over the last century in modeling the distribution of stars in elliptical galaxies, plus the bulges of lenticular and spiral galaxies and their surrounding disks. A number of nearly forgotten or poorly recognized references have been identified. The universality, or at least versatility, of Sérsic's $R^{1/n}$ model to describe bulges (➲ Sect. 4.1) and elliptical galaxies (➲ Sect. 2.1) extends to the stellar halos of cD galaxies (➲ Sect. 2.4) and simulated dark matter halos (➲ Sect. 2.1.1).

Dwarf and ordinary elliptical galaxies were shown in ➲ Sect. 3.1 to be united by two continuous linear relations between absolute magnitude and (a) the stellar concentration quantified through Sérsic's (1963) $R^{1/n}$ shape parameter (➲ Sect. 3.1.2) and (b) the central surface brightness, which is also related to the central density (➲ Sect. 3.1.1). As discussed in ➲ Sect. 3.3, a break in the latter relation at $M_B \approx -20.5$ signals the onset of partially depleted cores relative to the outer Sérsic profile in luminous elliptical galaxies. Additional scaling relations are also noted to show a change in character at this magnitude, which may denote the onset of dry galaxy merging.

The identification of depleted galaxy cores and excess nuclear light relative to the outer Sérsic profile was discussed in ➲ Sects. 2.2, ➲ 2.2.1, and ➲ 2.3. After accounting for these features, the above two linear relations result in curved scaling relations involving effective half light radii and effective surface brightness (➲ Sect. 3.2). Specifically, the M–R_{e}, M–μ_{e}, M–$\langle\mu\rangle_{\mathrm{e}}$, μ_{e}–R_{e}, $\langle\mu\rangle_{\mathrm{e}}$–$R_{\mathrm{e}}$, and n–R_{e} relations are nonlinear. These continuous curved relations exist because elliptical galaxies do not have a universal profile shape, such as an $R^{1/4}$ profile but instead a range of profile shapes that vary smoothly with absolute magnitude. Without an appreciation of the origin of these curved relations, they had in the past been heralded as evidence for a dichotomy between faint and bright elliptical galaxies. Numerical simulations and semi-numerical models which try to reproduce the full elliptical galaxy sequence must be able to reproduce these nonlinear relations. This will likely require physical processes which work in tandem, albeit to different degrees over different mass ranges, to produce a continuum of galaxy properties that scale with mass while adhering to the linear M–n and M–μ_0 relations (subject to core formation).

The structure of disk galaxies was reviewed in Sect. 4, with Sect. 4.1 covering the eventual recognition that bulges, like elliptical galaxies, are well described by the Sérsic model. A discussion of the difficulties in identifying pseudobulges was provided in Sect. 4.3, covering the shape of bulge light profiles, rotation, stellar ages, and nonlinear scaling relations. Additional sections briefly encompassed issues related to dust (Sect. 4.2), bulgeless galaxies (Sect. 4.4) and models for barred galaxies (Sect. 4.5).

The upcoming 2.6-m VLT Survey Telescope (VST, Arnaboldi et al. 1998; Capaccioli et al. 2005), plus the 4×1.8-m Pan-STARRS array (Kaiser et al. 2002), the 4-m Visible and Infrared Survey Telescope for Astronomy (VISTA, Emerson et al. 2004) and the 8.4-m Large Synoptic Survey Telescope (LSST, Tyson 2002) are expected to deliver sub-arcsecond, deep and wide field-of-view imaging covering thousands of resolvable galaxies. By pushing down the luminosity function into the dwarf galaxy regime, and through the application of improved galaxy parameterization methods which allow for structural nonhomology and the two- or three-component nature of disk galaxies, *both* statistical and systematic errors will be reduced. This will undoubtedly provide improved constraints on galaxy scaling relations and, in turn, a fuller understanding of galaxy evolution.

Acknowledgments

This research was supported under the Australian Research Council's Discovery Projects funding scheme (DP110103509). This research has made use of NASA's Astrophysics Data System (ADS) Bibliographic Services and the NASA/IPAC Extragalactic Database (NED).

References

Aceves, H., Velázquez, H., & Cruz, F. 2006, MNRAS, 373, 632

Ade, P. A. R., et al. 2011, (arXiv:1101.2045)

Aguerri, J. A. L., Beckman, J. E., & Prieto, M. 1998, AJ, 116, 2136

Aguerri, J. A. L., Balcells, M., & Peletier, R. F. 2001, A&A, 367, 428

Allen, P. D., Driver, S. P., Graham, A., Cameron, E., Liske, J., & De Propris, R. 2006, MNRAS, 371, 2

Andredakis, Y. C., & Sanders, R. H. 1994, MNRAS, 267, 283

Andredakis, Y. C., Peletier, R. F., & Balcells, M. 1995, MNRAS, 275, 874

Arnaboldi, M., Capaccioli, M., Mancini, D., Rafanelli, P., Scaramella, R., Sedmak G., & Vettolani G.P. 1998, The Messenger, 93, 30

Athanassoula, E. 1992, MNRAS, 259, 345

Athanassoula, E. 2005, MNRAS, 358, 1477

Athanassoula, E., et al. 1990, MNRAS, 245, 130

Auger, M. W., Treu, T., Bolton, A. S., Gavazzi, R., Koopmans, L. V. E., Marshall, P. J., Moustakas, L. A., & Burles, S. 2010, ApJ, 724, 511

Avila-Reese, V., Firmani, C., Klypin A., Kravtsov, A. V. 1999, MNRAS, 310, 527

Babcock, H. W. 1938, PASP, 50, 174

Babcock, H. W. 1939, Lick Obs Bull, 19, 41

Babusiaux, C., et al. 2010, A&A, 519, A77

Baes, M., & Van Hese, E. 2011, A&A, 534, A69

Balcells, M., Graham, A. W., Dominguez-Palmero, L., & Peletier, R. F. 2003, ApJ, 582, L79

Balcells, M., Graham, A. W., & Peletier, R. F. 2007, ApJ, 665, 1084

Baldry, I. K., Glazebrook, K., Brinkmann, J., Ivezić, Ž., Lupton, R. H., Nichol, R. C., & Szalay, A. S. 2004, ApJ, 600, 681

Barazza, F. D., Jogee, S., & Marinova, I. 2008, ApJ, 675, 1194

Bardeen, J. M. 1975, IAU Symp, 69, 297

Barnes, J. H. 1988, ApJ, 331, 699

Barth, A. J., Strigari, L. E., Bentz, M. C., Greene, J. E., & Ho, L. C. 2009, ApJ, 690, 1031

Barton, I. J., & Thompson, L. A. 1997, AJ, 114, 655

Bassino, L. P., Muzzio, J. C., & Rabolli, M. 1994, ApJ, 431, 634

Begelman, M. C., Blandford, R. D., & Rees, M. J. 1980, Nature, 287, 307

Bekki, K. 2010, MNRAS, 401, L58

Bekki, K., Couch, W. J., & Drinkwater, M. J. 2001a, ApJ, 552, L105

Bekki, K., Couch, W. J., Drinkwater, M. J., & Gregg, M. D. 2001b, ApJ, 557, L39

Bell, E. F., & de Jong, R. S. 2000, MNRAS, 312, 497

Bender, R., Doebereiner, S., & Moellenhoff, C. 1988, A&AS, 74, 385

Bernardi, M., et al. 2007, AJ, 133, 1741

Bertin, G., & Stiavelli, M. 1993, Rep Prog Phys, 56, 493

Bertin, G., Ciotti, L., & Del Principe, M. 2002, A&A, 386, 149

Bertola, F., & Capaccioli, M. 1977, ApJ, 211, 697

Bianchi, S. 2008, A&A, 490, 461

Binggeli, B., & Cameron, L. M. 1991, A&A, 252, 27

Binggeli, B., & Cameron, L. M. 1993, A&AS, 98, 297

Binggeli, B., & Jerjen, H. 1998, A&A, 333, 17

Binggeli, B., Sandage, A., & Tarenghi, M. 1984, AJ, 89, 64

Binggeli, B., Sandage, A., & Tammann, G. A. 1985, AJ, 90, 1681

Binney, J. 1978, MNRAS, 183, 501

Binney, J. 1982, ARA&A, 20, 399

Binney, J., & Mamon, G. A. 1982, MNRAS, 200, 361

Binney, J., Petrou, M. 1985, MNRAS, 214, 449

Birnboim, Y., & Dekel, A. 2003, MNRAS, 345, 349

Bland-Hawthorn, J., Vlajić, M., Freeman, K. C., & Draine, B. T. 2005, ApJ, 629, 239

Blanton, M., et al. 2005a, AJ, 129, 2562

Blanton, M. R., et al., 2005b, ApJ, 631, 208

Block, D. L., et al. 2004, in Astrophys. Space Sci. Libr. (ASSL) 319, Penetrating Bars Through Masks of Cosmic Dust, ed. D. Block, I. Puerari, K. C. Freeman, R. Groess, & E. K. Block (Dordrecht: Kluwer), 15

Böker, T., Stanek, R., & van der Marel, R. P. 2003, AJ, 125, 1073

Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., & Moustakas, L. A. 2006, ApJ, 638, 703

Boselli, A., Boissier, S., Cortese, L., Gavazzi, G. 2008, A&A, 489, 1015

Bothun, G. D., & Gregg, M. D. 1990, ApJ, 350, 73

Bothun, G. D., Mould, J. R., Caldwell, N., & MacGillivray, H. T. 1986, AJ, 92, 1007

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., Lacey, C. G. 2006, MNRAS, 370, 645

Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2004, ApJ, 613, L37

Bregman, J., Snider, B., Grego, L., & Cox, C. 1998, ApJ, 499, 670

Brodie, J. P., & Strader, J. 2006, ARA&A, 44, 193

Brook, C. B., et al. 2011, MNRAS, 415, 1051

Brown, R. J. N., et al. 2003, MNRAS, 341, 747

Buote, D. A., & Humphrey, P. J. 2011, in Astrophys. Space Sci. Libr. (ASSL), Hot Interstellar Matter in Elliptical Galaxies, ed. D.-W. Kim, & S. Pellegrini (Heidelberg/New York: Springer) (arXiv:1104.0012)

Burkert, A. 1995, ApJL, 447, L25

Byun, Y.-I., et al. 1996, AJ, 111, 1889

Caldwell, N. 1983, AJ, 88, 804

Caldwell, N., & Bothun, G. D. 1987, AJ, 94, 1126

Cameron, E., Driver, S. P., Graham, A. W., & Liske, J. 2009, ApJ, 699, 105

Caon, N., Capaccioli, M., & Rampazzo, R. 1990, A&AS, 86, 429

Caon, N., Capaccioli, M., & D'Onofrio, M. 1993, MNRAS, 265, 1013

Caon, N., Capaccioli, M., & D'Onofrio, M. 1994, A&AS, 106, 199

Capaccioli, M. 1985, in New Aspects of Galaxy Photometry, ed. J.-L. Nieto (Berlin/New York: Springer), 53

Capaccioli, M. 1987, in IAU Symp. 127, Structure and Dynamics of Elliptical Galaxies (Dordrecht: Reidel), 47

Capaccioli, M. 1989, in The World of Galaxies, ed. H. G. Corwin, & L. Bottinelli (Berlin: Springer), 208

Capaccioli, M., & Caon, N. 1991, MNRAS, 248, 523

Capaccioli, M., Mancini D., & Sedmak, G. 2005, The Messenger, 120, 10

Cardone, V. F., Piedipalumbo, E., & Tortora, C. 2005, MNRAS, 358, 1325

Carlberg, R. G., Lake, G., & Norman, C. A. 1986, ApJ, 300, L1

Carollo, C. M., Stiavelli, M., & Mack, J. 1998, AJ, 116, 68

Carollo, C. M., Scarlata, C., Stiavelli, M., Wyse, R. F. G., & Mayer, L. 2007, ApJ, 658, 960

Carter, D. 1977, MNRAS, 178, 137

Carter, D. 1978, MNRAS, 182, 797

Carter, D. 1987, ApJ, 312, 514

Carter, D., Pass, S., Kennedy, J., Karick, A. M., & Smith, R. J. 2011, MNRAS, 414, 3410

Cavichia, O., Costa, R. D. D., Mollá, M., & Maciel, W. J. 2012, in Planetary Nebulae: An Eye to the Future, IAU Symposium, 283, 326

Cecil, G., & Rose, J. A. 2007, Rep Prog Phys, 70, 1177

Cellone, S. A., Forte, J. C., & Geisler, D. 1994, ApJS, 93, 397

Chang, C.-K., Ko, C.-M., & Peng, T.-H. 2011, ApJ, in press (arXiv:1107.3884)

Chen, C.-W., Côté, P., West, A. A., Peng, E. W., & Ferrarese, L. 2010, ApJS, 191, 1

Chilingarian, I. 2009, MNRAS, 394, 1229

Chilingarian, I., Cayatte, V., Revaz, Y., Dodonov, S., Durand, D., Durret, F., Micol, A., & Slezak, E. 2009, Science, 326, 1379

Ciotti, L. 1991, A&A, 249, 99

Ciotti, L. 2009, Nuovo Cimento Riv Ser, 32, 1

Ciotti, L., & Bertin, G. 1999, A&A, 352, 447

Clemens, M. S., et al. 2010, A&A, 518, L50

Cole, S., Lacey, C. G., Baugh, C. M., & Frenk, C. S. 2000, MNRAS, 319, 168

Combes, F., & Sanders, R. H., 1981, A&A, 96, 164

Combes, F., Debbasch, F., Friedli, D., & Pfenniger, D. 1990, A&A, 233, 82

Conselice, C. J., Bluck, A. F. L., Ravindranath, S., Mortlock, A., Koekemoer, A., Buitrago, F., Grütbauch, R., & Penny, S. 2011, MNRAS, submitted (arXiv:1105.2522)

Côté, P., et al. 2006, ApJ, 165, 57

Côté, P., et al. 2007, ApJ, 671, 1456

Côté, P., et al. 2008, in Dynamical Evolution of Dense Stellar Systems, Proceedings of the International Astronomical Union, IAU Symposium. eds: E. Vesperini, M. Giersz & A. Sills (Cambridge University Press), 246, 377–386

Covington, M. D., Primack, J. R., Porter, L. A., Croton, D. J., Somerville, R. S., & Dekel A. 2011, MNRAS, 415, 3135

Crane, P., et al. 1993, AJ, 106, 1371

Croton, D. J., et al. 2006, MNRAS, 365, 11

Curtis, H. D. 1918, Lick Obs Pub, 13, 9

Dabringhausen, J., Hilker, M., & Kroupa, P. 2008, MNRAS, 386, 864

Daddi, E., et al. 2005, ApJ, 626, 680

Damjanov, I., et al. 2009, ApJ, 695, 101

Davidge, T. J., Beck, T. L., & McGregor, P. J. 2008, ApJ, 677, 238

Davies, R. L., Efstathiou, G., Fall, S. M., Illingworth, G., & Schechter, P. L. 1983, ApJ, 266, 41

Davies, J. I., Phillipps, S., Cawson, M. G. M., Disney, M. J., & Kibblewhite, E. J. 1988, MNRAS, 232, 239

de Grijs, R., Peletier, R. F., & van der Kruit, P. C. 1997, A&A, 327, 966

de Jong, R. 1996, A&AS, 118, 557

de Jong, R. S., & van der Kruit, P. C. 1994, A&AS, 106, 451

De Lucia, G., Springel, V., White, S. D. M., Croton, D., Kauffmann, G. 2006, MNRAS, 366, 499

de Rijcke, S., Michielsen, D., Dejonghe, H., Zeilinger, W. W., & Hau, G. K. T. 2005, A&A, 438, 491

Del Popolo, A. 2010, MNRAS, 408, 1808

de Sitter, W. 1917, MNRAS, 78, 3

de Vaucouleurs, G. 1948, Ann d'Astrophys, 11, 247

de Vaucouleurs, G. 1953, MNRAS, 113, 134

de Vaucouleurs, G. 1957, AJ, 62, 69

de Vaucouleurs, G. 1958, ApJ, 128, 465

de Vaucouleurs, G. 1959, in Handbuch der Physik, ed. S. Flügge (Berlin: Springer), 311

de Vaucouleurs, G. 1961, ApJS, 5, 233

de Vaucouleurs, G. 1963, ApJS, 8, 31

de Vaucouleurs, G. 1969, ApJL, 4, L17

de Vaucouleurs, G. 1974, IAUS, 58, 1

de Vaucouleurs, G., & Capaccioli, M. 1979, ApJS, 40, 669

de Vaucouleurs, G., & de Vaucouleurs, A. 1970, ApJL, 5, L219

Devereux, N. A., & Young, J. S. 1990, ApJ, 359, 42

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., Jr., Buta, R. J., Paturel, G., & Fouque, P. 1991, Third Reference Catalog (Berlin/Heidelberg/New York: Springer)

de Zeeuw, P. T., & Franx, M. 1991, ARA&A, 29, 239

Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604

Djorgovski, S., & Davis, M. 1987, ApJ, 313, 59

Domínguez-Tenreiro, R., Tissera, P. B., & Sáiz, A. 1998, Ap&SS, 263, 35

Donato, F., et al. 2009, MNRAS, 397, 1169

Dong, X. Y., & De Robertis, M. M. 2006, AJ, 131, 1236

D'Onghia, E., Burkert, A., Murante, G., & Khochfar, S. 2006, MNRAS, 372, 1525

D'Onofrio, M. 2001, MNRAS, 326, 1517

D'Onofrio, M., Capaccioli, M., & Caon, N. 1994, MNRAS, 271, 523

Draine, B. T. 2003, ARA&A, 41, 241

Dreyer, J. L. E. 1888, Mem R Astron Soc, 49, 1

Dreyer, J. L. E. 1895, Mem R Astron Soc, 51, 185

Dreyer, J. L. E. 1908, Mem R Astron Soc, 59, 105

Drinkwater, M. J., Jones, J. B., Gregg, M. D., & Phillipps, S. 2000, PASA, 17, 227

Driver, S. P. 2010, AIP Conf Ser, 1240, 17 (arXiv:1001.4054)

Driver, S. P., Allen, P. D., Liske, J., & Graham, A. 2007a, MNRAS, 379, 1022

Driver, S. P., et al. 2007b, ApJ, 657, L85

Driver, S. P., Popescu, C. C., Tuffs, R. J., Graham, A. W., Liske, J., & Baldry, I. 2008, ApJ, 678, L101

Driver, S. P. 2010, Amer. Inst. Phys. Conf. Ser., 1240, 17

Dullo, B. T., & Graham, A. W. 2012, ApJ, 755, 163

Duncan, M. J., & Wheeler, J. C. 1980, ApJ, 237, L27

Ebisuzaki, T., Makino, J., & Okumura, S. K. 1991, Nature, 354, 212

Jaan Einasto, J. 1965, Trudy Inst Astrofiz Alma-Ata, 5, 87

Ellis, S. C., & O'Sullivan, E. 2006, MNRAS, 367, 627

Elmegreen, B. G., Elmegreen, D. M., Chromey, F. R., Hasselbacher, D. A., & Bissell, B. A. 1996, AJ, 111, 2233

Elson, R. A. W., Fall, S. M., & Freeman, K. C. 1987, ApJ, 323, 54

Emerson, J. P., Sutherland, W. J., McPherson, A. M., Craig, S. C., Dalton, G. B., & Ward, A. K. 2004, The Messenger, 117, 27

Emsellem, E., et al. 2007, MNRAS, 379, 401

Emsellem, E., et al. 2011, MNRAS, in press (arXiv:1102.4444)

Erwin, P., Beltrán, J. C. V., Graham, A. W., & Beckman, J. E. 2003, ApJ, 597, 929

Erwin, P., Beckman, J. E., & Pohlen, M. 2005, ApJ, 626, L81

Eskridge, P. B., et al. 2000, AJ, 119, 536

Evstigneeva, E. A., Gregg, M. D., Drinkwater, M. J., & Hilker, M. 2007, AJ, 133, 1722

Faber, S. M., & Jackson, R. E. 1976, ApJ, 204, 668

Faber, S. M., & Lin, D. M. C. 1983, ApJ, 266, L17

Faber, S. M., et al. 1997, AJ, 114, 1771

Farouki, R. T., Shapiro, S. L., & Duncan, M. J. 1983, ApJ, 265, 597

Ferguson, H. C., & Binggeli, B. 1994, A&ARv, 6, 67

Ferrarese, L., & Ford, H. 2005, Space Sci Rev, 116, 523

Ferrarese, L., van den Bosch, F. C., Ford, H. C., Jaffe, W., & O'Connell, R. W. 1994, AJ, 108, 1598

Ferrarese, L., et al. 2006a, ApJS, 164, 334

Ferrarese, L., et al. 2006b, (arXiv:astro-ph/0612139)

Ferrari, F., Dottori, H., Caon, N., Nobrega, A., & Pavani, D. B. 2004, MNRAS, 347, 824

Ferrers, N. M. 1877, Q. J Pure Appl Math, 14, 1

Filippenko, A. V., & Ho, L. C. 2003, ApJ, 588, L13

Fisher, D. B., & Drory, N. 2008, AJ, 136, 773

Fisher, D. B., & Drory, N. 2010, ApJ, 716, 942

Fisher, D., Franx, M., & Illingworth, G. 1996, ApJ, 459, 110

Fontanot, F., De Lucia, G., Wilman, D., & Monaco, P. 2011, MNRAS, in press (arXiv:1102.3188)

Forbes, D. A., & Kroupa, P. 2011, PASA, 28, 77

Forbes, D. A., Franx, M., & Illingworth, G. D. 1994, ApJ, 428, L49

Forbes, D. A., Lasky, P., Graham A. W., & Spitler, L. 2008, MNRAS, 389, 1924

Frankston, M., & Schild, R. 1976, AJ, 81, 500

Freeman, K. C. 1966, MNRAS, 133, 47

Freeman, K. C. 1970, ApJ, 160, 811

Freeman, K. C. 1990, in Dynamics and Interactions of Galaxies, ed. R. Wielen (Berlin: Springer), 36

Freeman, K. C. 2004, in Astrophys. Space Sci. Libr. (ASSL) 319, Penetrating Bars Through Masks of Cosmic Dust – The Hubble Tuning Fork strikes a New Note, ed. D. Block, I. Puerari, K. C. Freeman, R. Groess, & E. K. Block (Dordrecht: Kluwer), 639

Friedmann, A. 1922, Uner die Krümmung des Raumes. Z Phys, x, 377–86

Gadotti, D. A. 2008, MNRAS, 384, 420

Gadotti, D. A., 2009, MNRAS, 393, 1531

Gadotti, D. A. 2011, MNRAS, submitted (arXiv:1101.2900)

Gadotti, D. A., & dos Anjos, S. 2001, AJ, 122, 1298

Gavazzi, G., Donati, A., Cucciati, O., Sabatini, S., Boselli, A., Davies, J., & Zibetti, S. 2005, A&A, 430, 411

Gavazzi, R., Treu, T., Rhodes, J. D., Koopmans, L. V. E., Bolton, A. S., Burles, S., Massey, R. J., & Moustakas, L. A. 2007, ApJ, 667, 176

Gebhardt, K., et al. 2001, AJ, 122, 2469

Geha, M., Guhathakurta, P., & van der Marel, R. P. 2003, AJ, 126, 1794

Genzel, R., Thatte, N., Krabbe, A., Kroker, H., & Tacconi-Garman, L. E. 1996, ApJ, 472, 153

Gerbal, D., Lima Neto, G. B., Márquez, I., & Verhagen, H. 1997, MNRAS, 285, L41

Gil de Paz, A., et al. 2007, ApJS, 173, 185

Gilmore, G., & Reid, N. 1983, MNRAS, 202, 1025

Gilmore, G., Wilkinson, M. I., Wyse, R. F. G., Kleyna, J. T., Koch, A., Evans, N. W., & Grebel, E. K. 2007, ApJ, 663, 948

Glass, L., et al. 2011, ApJ, 726, 31

Gliozzi, M., Satyapal, S., Eracleous, M., Titarchuk, L., & Cheung, C. C. 2009, ApJ, 700, 1759

Gonzalez, A. H., Zabludoff, A. I., & Zaritsky, D. 2005, ApJ, 618, 195

González-García, A. C., Oñorbe, J., Domínguez-Tenreiro, R., & Gómez-Flechoso, M. Á. 2009, A&A, 497, 35

Goudfrooij, P., & de Jong, T. 1995, A&A, 298, 784

Goudfrooij, P., Gorgas, J., & Jablonka, P. 1999, Ap&SS, 269, 109

Governato, F., et al. 2010, Nature, 463, 203

Graham, A. W. 2001, AJ, 121, 820

Graham, A. W. 2002, ApJ, 568, L13

Graham, A. W. 2004, ApJ, 613, L33

Graham, A. W. 2005, in IAU Colloc. 198, Near-Field Cosmology with Dwarf Elliptical Galaxies, ed. H. Jerjen, & B. Binggeli (Cambridge: Cambridge University Press), 303

Graham, A. W. 2007, MNRAS, 379, 711

Graham, A. W. 2010, in A Universe of Dwarf Galaxies, ed. F. Prugniel (arXiv:1009.5002)

Graham, A. W. 2011, arXiv:1103.0525

Graham, A. W., & Driver S. P. 2005, PASA, 22(2), 118

Graham, A. W., & Driver, S. P. 2007, ApJ, 655, 77

Graham, A. W., & Guzmán, R. 2003, AJ, 125, 2936

Graham, A. W., Guzmán, R. 2004, in Penetrating Bars Through Masks of Cosmic Dust, ed. D. L. Block et al. (Dordrecht: Kluwer), 723

Graham, A. W., & Spitler L. R. 2009, MNRAS, 397, 2148

Graham, A. W., & Worley, C. C. 2008, MNRAS, 388, 1708

Graham, A. W., Lauer, T. R., Colless, M. M., & Postman, M. 1996, ApJ, 465, 534

Graham, A. W., Colless, M. M., Busarello, G., Zaggia S., & Longo, G. 1998, A&AS, 133, 325

Graham, A. W., Trujillo, I., & Caon, N. 2001, AJ, 122, 1707

Graham, A. W., Erwin, P., Trujillo, I., & Asensio Ramos, A. 2003, AJ, 125, 2951

Graham, A. W., Driver, S. P., Petrosian, V., Conselice, C. J., Bershady, M. A., Crawford, S. M., & Goto, T. 2005, AJ, 130, 1535

Graham, A. W., Merritt, D., Moore, B., Diemand, J., & Terzić, B. 2006, AJ, 132, 2701

Grant, N. I., Kuipers, J. A., & Phillipps, S. 2005, MNRAS, 363, 1019

Grebel, E. K. 2001, ApSSS, 277, 231

Greene, J. E., Ho, L. C., & Barth, A. J. 2008, ApJ, 688, 159

Grillmair, C. J., Faber, S. M., Lauer, T. R., Baum, W. A., Lynds, R. C., O'Neil, E. J., Jr., & Shaya, E. J. 1994, AJ, 108, 102

Guo, Q., et al. 2011, MNRAS, 413, 101

Haehnelt, M. G., & Kauffmann, H. 2002, MNRAS, 336, L61

Harris, W. E., & van den Bergh, S. 1981, AJ, 86, 1627

Haşegan, M., et al. 2005, ApJ, 627, 203

Hawarden, T. G., Longmore, A. J., Tritton, S. B., Elson, R. A. W., & Corwin, H. G., Jr. 1981, MNRAS, 196, 747

Held, E. V., de Zeeuw, T., Mould, J., & Picard, A. 1992, AJ, 103, 851

Heraudeau, P., & Simien, F. 1997, A&A, 326, 897

Hernquist, L. 1990, ApJ, 356, 359

Herschel, J. 1864, Catalogue of Nebulae and clusters of stars. R Soc Lond Philos Trans Ser I, 154, 1

Hilker, M., Infante, L., Vieira, G., Kissler-Patig, M., & Richtler T. 1999, A&AS, 134, 75

Hill, D. T., et al. 2011, MNRAS, 412, 765

Hjorth, J., & Madsen, J. 1995, ApJ, 445, 55

Hodge, P. W. 1961a, AJ, 66, 249

Hodge, P. W. 1961b, AJ, 66, 384

Hodge, P. W. 1963, AJ, 68, 691

Hodge, P. W. 1964, AJ, 69, 442

Hodge, P. W. 1971, ARA&A, 9, 35

Hoessel, J. G., Oegerle, W. R., & Schneider, D. P. 1987, AJ, 94, 1111

Hoffman, L., Cox, T. J., Dutta, S., & Hernquist, L. 2009, ApJ, 705, 920

Hohl, F. 1975, IAU Symp, 69, 349

Hohl, F., & Zhang, T. A. 1979, AJ, 84, 585

Hoskin, M. 1970, J Hist Astron, 1, 44

Hubble, E. 1926, ApJ, 64, 321

Hubble, E. 1929, Proc Natl Acad Sci, 15, 168

Hubble, E. 1930, ApJ, 71, 231

Hubble, E. 1934, The Halley Lecture, delivered on 8 May 1934 (Oxford: Clarendon Press, 1934)

Hubble, E. P. 1936a, Realm of the Nebulae, ed. E. P. Hubble (New Haven: Yale University Press)

Hubble, E. P. 1936b, ApJ, 84, 517

Hubble, E. P. 1937, MNRAS, 97, 513

Hubble, E., & Tolman, R., 1935, ApJ, 82, 302

Humasson, M. 1929, Proc Natl Acad Sci, 15, 167

Humphre, P. J., & Buote, D. A. 2010, MNRAS, 403, 2143

Huxor, A. P., Phillipps, S., Price, J., & Harniman R. 2011, MNRAS, in press (arXiv:1103.1257)

Ichikawa, S.-I., Wakamatsu, K.-I., & Okumura, S. K. 1986, ApJS, 60, 475

Iodice, E., D'Onofrio, M., & Capaccioli, M. 1997, ASP Conf Ser, 116, 841

Iodice, E., D'Onofrio, M., & Capaccioli, M. 1999, ASP Conf Ser, 176, 402

Ivezić, Ž. et al. 2004, Astron Nachr, 325, 583

Jablonka, P., Gorgas, J., & Goudfrooij, P. 2007, A&A, 474, 763

Jaffe, W., Ford, H. C., O'Connell, R. W., van den Bosch, F. C., & Ferrarese, L. 1994, AJ, 108, 1567

Janz, J., & Lisker, T. 2008, ApJ, 689, L25

Janz, J., & Lisker, T. 2009, ApJ, 696, 102

Jeans, J. 1919, Problems of Cosmogony and Stellar Dynamics (Cambridge: Cambridge University Press)

Jeans, J. H. 1928, Astronomy & Cosmogony (Cambridge: Cambridge University Press), 332

Jerjen, H., & Binggeli, B. 1997, in ASP Conf. Ser. 116, The Nature of Elliptical Galaxies; The Second Stromlo Symposium, ed. M. Arnaboldi, G. S. Da Costa, & P. Saha (San Francisco, CA: ASP), 239

Jerjen, H., Kalnajs, A., & Binggeli, B. 2000b, A&A, 358, 845

Jiménez, N., Cora, S. A., Bassino, L. P., Tecce, T. E., & Smith Castelli, A. V. 2011, MNRAS, in press (arXiv:1107.0722)

Jing, Y. P., & Suto, Y. 2000, ApJ, 529, L69

Jonsson, P. 2006, MNRAS, 372, 2

Jonsson, P., Groves, B. A., & Cox, T. J. 2010, MNRAS, 403, 17

Kaiser, N., et al. 2002, Proc SPIE, 4836, 154

Kannappan, S. J., Jansen, R. A., & Barton, E. J. 2004, AJ, 127, 1371

Kant, I. 1755, Allgemeine Naturgeschichte und Theorie des Himmels (Königsberg/Leipzig: Johann Friederich Petersen)

Karachentseva, V. E., Prugniel, P., Vennik, J., Richter, G. M., Thuan, T. X., & Martin, J. M. 1996, A&ASS, 117, 343

Kauffmann, G., White, S. D. M., Heckman, T. M., Ménard, B., Brinchmann, J., Charlot, S., Tremonti, C., & Brinkmann, J. 2004, MNRAS, 353, 713

Kautsch, S. J. 2009a, Astron Nachr, 330, 1053

Kautsch, S. J. 2009b, PASP, 121, 1297

Kawata, D., Cen, R., & Ho, L. C. 2007, ApJ, 669, 232

Keel, W. C., White, R. E., III, 2001a, AJ, 121, 1442

Keel, W. C., & White, R. E., III, 2001b, AJ, 122, 1369

Kent, S. 1985, ApJS, 59, 115

Kent, S. M., Dame, T., & Fazio, G., 1991, ApJ, 378, 131

Kereš, D., Katz, N., Weinberg, D. H., & Davé, R. 2005, MNRAS, 363, 2

Khosroshahi, H. G., Wadadekar, Y., & Kembhavi, A. 2000, ApJ, 533, 162

Khochfar, S., & Burkert, A. 2003, ApJ Letters, 597, L117

King, I. R. 1962, AJ, 67, 471

King, I. R. 1966, AJ, 71, 64

King, I. R. 1978, ApJ, 222, 1

King, I. R., & Minkowski, R. 1966, ApJ, 143, 1002

King, I. R., & Minkowski, R. 1972, IAU Symp, 44, 87

Kochanek, C. S. 1995, ApJ, 445, 559

Koda, J., Milosavljević, M., & Shapiro, P. R. 2009, ApJ, 696, 254

Kodaira, K., Watanabe, M., & Okamura, S. 1986, ApJS, 62, 703

Kormendy J., Freeman K. C. 2004, Dark Matter in Galaxies, IAU Symp., 220, 377

Knapen, J. H. 2010, in Galaxies and Their Masks, ed. D. L. Block, K. C. Freeman, & I. Puerari (New York: Springer), 201 (arXiv:1005.0506)

Knapen, J. H., de Jong, R. S., Stedman, S., & Bramich, D. M. 2003, MNRAS, 344, 527

Knox Shaw, H. 1915, Helwan Obs Bull No, 15, 129

Koopmans, L. V. E., Treu, T., Bolton, A. S., Burles, S., & Moustakas, L. A. 2006, ApJ, 649, 599

Koopmans, L. V. E., et al. 2009, ApJ, 703, L51

Kormendy, J. 1977a, ApJ, 217, 406

Kormendy, J. 1977b, ApJ, 218, 333

Kormendy, J. 1982, in Saas-Fee Advanced Course 12: Morphology and Dynamics of Galaxies (Sauverny: Geneva Observatory), 113

Kormendy, J. 1985a, ApJ, 292, L9

Kormendy, J. 1985b, ApJ, 295, 73

Kormendy, J., & Kennicutt, R. C., Jr. 2004, ARA&A, 42, 603

Kormendy, J., & Freeman, K. C. 2004, Dark Matter in Galaxies, IAU Symp., 220, 377

Kormendy, J., Dressler, A., Byun, Y. I., Faber, S. M., Grillmair, C., Lauer, T. R., Richstone, D., & Tremaine, S. 1994, in ESO Conf. Ser. 49, Dwarf Galaxies, ed. G. Meylan, & P. Prugniel (Garching: ESO), 147

Kormendy, J., Gebhardt, K., Fisher, D. B., Drory, N., Macchetto, F. D., & Sparks, W. B. 2005, AJ, 129, 2636

Kormendy, J., Fisher, D. B., Cornell, M. E., & Bender, R. 2009, ApJS, 182, 216

Kragh, H., & Smith, R. W. 2003, Hist Sci, 41, 141

Krajnović, D., et al. 2008, MNRAS, 390, 93

Larson, R. B. 1969, MNRAS, 145, 405

Larson, R. B. 1974, MNRAS, 166, 585

Larson, R. B. 1975, MNRAS, 173, 671

Lauer, T. R. 1983, in Elliptical Galaxies, Surface Photometry (Santa Cruz: University of California)

Lauer, T. R. 1984, Bull Am Astron Soc, 16, 455 (08.03)

Lauer, T. R. 1985, ApJ, 292, 104

Lauer, T. R., et al. 1995, AJ, 110, 2622

Lauer, T. R., et al. 2005, AJ, 129, 2138

Lauer, T. R., et al., 2007, ApJ, 662, 808

Launhardt, R., Zylka, R., & Mezger, P. G. 2002, A&A, 384, 112

Laurikainen, E., & Salo, H. 2002, MNRAS, 337, 1118

Laurikainen, E., Salo, H., & Buta, R. 2005, MNRAS, 362, 1319

Laurikainen, E., Salo, H., Buta, R., & Knapen, J. H. 2007, MNRAS, 381, 401

Lemaitre, G. 1927, Annales de la Societe Scientique de Bruxelles, 47, 49

Liller, M. H. 1960, ApJ, 132, 306

Liller, M. H. 1966, ApJ, 146, 28

Lima Neto, G. B., Gerbal, D., & Márquez, I. 1999, MNRAS, 309, 481

Liske, J., Lemon, D. J., Driver, S. P., Cross, N. J. G., & Couch, W. J. 2003, MNRAS, 344, 307

Lisker, T. 2009, AN, 330, 1403

Lisker, T., & Han, Z. 2008, ApJ, 680, 1042

Lisker, T., Grebel, E. K., & Binggeli, B. 2005, IAU Colloq. 198, Near-Fields Cosmology with Dwarf Elliptical Galaxies, ed. H. Jerjen, & B. Binggeli (Cambridge: Cambridge University Press), 311

Lisker, T., Grebel, E. K., & Binggeli, B. 2006a, AJ, 132, 497

Lisker, T., Glatt, K., Westera, P., & Grebel, E. K. 2006b, AJ, 132, 2432

Lisker, T., et al. 2007, ApJ, 660, 1186

Liu, F. S., Xia, X. Y., Mao, S., Wu, H., Deng, Z. G. 2008, MNRAS, 385, 23

Lugger, P. M. 1984, ApJ, 286, 106

Lundmark, K. 1924, MNRAS, 84, 747

Ma, J., et al. 2007, MNRAS, 376, 1621

MacArthur, L. A., Courteau, S., & Holtzman, J. A. 2003, ApJ, 582, 689

MacArthur, L. A., González, J. J., & Courteau, S. 2009, MNRAS, 395, 28

MacArthur, L. A., McDonald, M., Courteau, S., & Jesús González, J. 2010, ApJ, 718, 768

Magorrian, J., et al. 1998, AJ, 115, 2285

Maller, A. H., Berlind, A. A., Blanton, M. R., & Hogg, D. W. 2009, ApJ, 691, 394
Malumuth, E. M., & Kirshner, R. P. 1981, ApJ, 251, 508
Mamon, G. A., & Łokas, E. L. 2005, MNRAS, 362, 95
Mandelbaum, R., et al. 2005, MNRAS, 361, 1287
Marconi, A., & Hunt, L. K. 2003, ApJ, 589, L21
Marinova, I., et al. 2012, ApJ, 746, 136
Márquez, I., Lima Neto, G. B., Capelato, H., Durret, F., Lanzoni, B., & Gerbal, D. 2001, A&A, 379, 767
Martin, P. 1995, AJ, 109, 2428
Martizzi, D., Teyssier, R., & Moore, B. 2011, MNRAS, submitted (arXiv:1106.5371)
Masters, K. L., Giovanelli, R., & Haynes, M. P. 2003, AJ, 126, 158
Matković, A., & Guzmán, R. 2005, MNRAS, 362, 289
Matthews, T. A., Morgan, W. W., & Schmidt, M. 1964, ApJ, 140, 35
McAlpine, W., Satyapal, S., Gliozzi, M., Cheung, C. C., Sambruna, R. M., & Eracleous, M. 2011, ApJ, 728, 25
McGlynn, T. A. 1984, ApJ, 281, 13
McLure, R. J., & Dunlop, J. S. 2004, MNRAS, 352, 1390
Mebold, U., Goss, W. M., Siegman, B., van Woerden, H., & Hawarden, T. G. 1979, A&A, 74, 100
Merritt, D. 2006a, ApJ, 648, 976
Merritt, D. 2006b, Rep Prog Phys, 69, 2513
Merritt, D., & Milosavljević, M. 2005, Living Rev Relativ, 8, 8
Merritt, D., Ferrarese L., & Joseph, C. L. 2001, Science, 293, 1116
Merritt, D., Navarro, J. F., Ludlow, A., & Jenkins, A. 2005, ApJL, 624, L85
Merritt, D., Graham, A. W., Moore, B., Diemand, J., & Terzić, B. 2006, AJ, 132, 2685
Metcalfe, N., Godwin, J. G., & Peach, J. V. 1994, MNRAS, 267, 431
Michard, R. 1979, A&A, 74, 206
Michard, R. 1985, A&AS, 59, 205
Mihos, J. C., & Hernquist, L. 1994. ApJ, 425, L13
Milosavljević, M., & Merritt, D. 2001, ApJ, 563, 34
Minkowski, R. 1962, IAU Symp, 15 112
Minniti, D., Saito, R. K., Alonso-Garcia, J., Lucas, P. W., & Hempel, M. 2011, (arXiv:1105.3151)
Misgeld, I., & Hilker, M. 2011, MNRAS, in press (arXiv:1103.1628)
Misgeld, I., Mieske, S., & Hilker, M. 2008, A&A, 486, 697
Misgeld, I., Hilker, M., & Mieske, S. 2009, A&A, 496, 683
Möllenhoff, C., & Heidt, J. 2001, A&A, 368, 16
Morgan, W. W., & Lesh, J. R., 1965, ApJ, 142, 1364
Morton, D. C., & Chevalier, R. A. 1973, ApJ, 179, 55
Moorthy, B. K., & Holtzman, J. A. 2006, MNRAS, 371, 583
Naab, T., Burkert, A., & Hernquist, L. 1999, ApJ, 523, L133
Naab, T., Khochfar, S., & Burkert, A. 2006, ApJ, 636, L81
Nair P., van den Bergh S., Abraham R. G. 2011, ApJ, 734, L31
Navarro, J. F., & Benz, W. 1991, ApJ, 380, 320
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Navarro, J. F., et al. 2004, MNRAS, 349, 1039
Nieto, J.-L. 1990, in Dynamics and Interactions of Galaxies, ed. R. Wielen (Berlin: Springer), 258
Nipoti, C., Londrillo, P., & Ciotti, L. 2006, MNRAS, 370, 681
Nipoti, C., Treu, T., Auger, M. W., & Bolton, A. S. 2009, ApJ, 706, L86
Norman, C. A., Sellwood, J. A., & Hasan, H. 1996, ApJ, 462, 114
Norris, M. A., & Kannappan, S. J. 2011, MNRAS, 414, 739
Oemler, A., Jr. 1976, ApJ, 209, 693
Ohta, K., Hamabe, M., & Wakamatsu, K.-I. 1990, ApJ, 357, 71
Ostriker, J. P. 1977, Proc Natl Acad Sci, 74, 1767
Parson, L. 1878, Sci Trans R Dublin Soc, II
Patterson, F. S. 1940, Harv Coll Obs Bull, 914, 9
Pease, F.G. 1918, Proceedings of the National Academy of Science, 4, 21
Peletier, R. F., & Balcells, M. 1996a, AJ, 111, 2238
Peletier, R. F., & Balcells, M. 1996b, IAUS, 171, 29
Peletier, R. F., Davies, R. L., Illingworth, G. D., Davis, L. E., & Cawson, M. 1990, AJ, 100, 1091
Peletier, R. F., Balcells, M., Davies, R. L., Andredakis, Y., Vazdekis, A., Burkert, A., & Prada, F. 1999, MNRAS, 310, 703
Peletier, R. F., et al. 2007, MNRAS, 379, 445
Pellegrini, S. 2010, ApJ, 717, 640
Pellet, A. 1976, A&A, 50, 421
Penny, S. J., & Conselice, C. J. 2008, MNRAS, 383, 247
Persic M., Salucci P., Stel F. 1996, MNRAS, 281, 27
Peterson, C. J. 1978, ApJ, 221, 80
Petrosian, V. 1976, ApJ, 209, L1
Pfenniger, D. 1993, in Galactic Bulges, ed. H. Dejonghe, & H. J. Habing (Dordrecht: Kluwer), 387
Pichon, C., Pogosyan, D., Kimm, T., Slyz, A., Devriendt, J., & Dubois, Y. 2011 (arXiv:1105.0210)
Pierini, D., Zibetti, S., Braglia, F., Böhringer, H., Finoguenov, A., Lynam, P. D., & Zhang, Y.-Y. 2008, A&A, 483, 727
Pilbratt, G. L., et al. 2010, A&A, 518, L1
Plummer, H. C. 1911, MNRAS, 71, 460
Pohlen, M., Beckman, J. E., Hüttemeister, S. H., Knapen, J. H., Erwin, P., & Dettmar, R.-J. 2004, in Penetrating Bars Through Masks of Cosmic

Dust, ed. D. L. Block et al. (Dordrecht: Kluwer), 713

Popescu, C. C., Misiriotis, A., Kylafis, N. D., Tuffs, R. J., & Fischera, J. 2000, A&A, 362, 138

Popescu, C. C., Tuffs, R. J., Dopita, M. A., Fischera, J., Kylafis, N. D., & Madore, B. F. 2011, A&A, 527, A109

Price, J. et al. 2009, MNRAS, 397, 1816

Prieto, M., Gottesman, S. T., Aguerri, J.-A. L., & Varela, A.-M. 1997, AJ, 114, 1413

Prieto, M., Aguerri, J. A. L., Varela, A. M., & Muñoz-Tuñón, C. 2001, A&A, 367, 405

Prugniel, P., & Simien, F. 1997, A&A, 321, 111

Qu, Y., Di Matteo, P., Lehnert, M. D., & van Driel, W. 2011, A&A, 530, A10

Ravindranath, S., Ho, L. C., Peng, C. Y., Filippenko, A. V., & Sargent, W. L. W. 2001, AJ, 122, 653

Reaves, G. 1956, AJ, 61, 69

Reaves, G. 1977, ApJS, 53, 375

Reese, A. S., Williams, T. B., Sellwood, J. A., Barnes, E. I., & Powell, B. A. 2007, AJ, 133, 2846

Renzini, A. 1999, in The Formation of Galactic Bulges, ed. C. M. Carollo, H. C. Ferguson, & R. F. G. Wyse (Cambridge: Cambridge University Press), 1

Renzini, A. 2006, ARA&A, 44, 141

Rest, A., et al. 2001, AJ, 121, 2431

Reynolds, J. H. 1913, MNRAS, 74, 132

Reynolds, J. H. 1920, Observatory, 43, 377

Reynolds, J. H. 1927, Observatory, 50, 185

Rich, R. M., & Origlia L. 2005, ApJ, 634, 1293

Richings, A. K., Uttley, P., & Kr̈ding, E. 2011, MNRAS, in press (arXiv:1104.1053)

Rix, H.-W., & White, S. D. M. 1990, ApJ, 362, 52

Robertson H. P. 1928, Phil. Mag., 5, 835

Robotham, A. S. G., & Driver, S. P. 2011, MNRAS, 413, 2570

Romanishin, W., Strom, K. M., & Strom, S. E. 1977, Bull Am Astron Soc, 9, 347

Rood, H. J., Page, T. L., Kintner, E. C., & King, I. R. 1972, ApJ, 175, 627

Rozas, M., Knapen, J. H., & Beckman, J. E. 1998, MNRAS, 301, 631

Rubin, V. C., Ford, W. K., Krishna Kumar, C. 1973, ApJ, 181, 61

Ryś A., & Falcón-Barroso J., 2010, in Dwarf Galaxies: Keys to Galaxy Formation and Evolution, Astrophysics and Space Science Proceedings, ISBN 978-3-642-22017-3. (Springer-Verlag Berlin Heidelberg), 2012, p. 155

Saha, K., Martinez-Valpuesta, I., & Gerhard, O. 2011, MNRAS, submitted (arXiv:1105.5797)

Sandage, A. 1972, ApJ, 176, 21

Sandage, A. 2004, in Astrophys. Space Sci. Libr. (ASSL) 319, Penetrating Bars Through Masks of Cosmic Dust – the Hubble Tuning Fork strikes a New Note, ed. D. Block et al. (Dordrecht: Kluwer), 39

Sandage, A. 2005, ARA&A, 43, 581

Sandage, A., & Binggeli, B. 1984, AJ, 89, 919

Saslaw, W. C., Valtonen, M. J., & Aarseth, S. J. 1974, ApJ, 190, 253

Scannapieco, C., White, S. D. M., Springel, V., & Tissera, P. B. 2011, MNRAS, 417, 154)

Schechter, P. L. 1980 AJ, 85, 801

Schödel, R., Merritt, D., & Eckart, A. 2009, J Phys Conf Ser, 131, 012044

Schombert, J. M. 1986, ApJS, 60, 603

Schurer, A., Calura, F., Silva, L., Pipino, A., Granato, G. L., Matteucci, F., & Maiolino, R. 2009, MNRAS, 394, 2001

Schwarzkopf, U., & Dettmar, R.-J. 1997, AGM, 13, 238

Seares, F. H. 1928, PASP, 40, 303

Seigar, M. S., & James, P. A. 1998, MNRAS, 299, 672

Seigar, M. S., Graham, A. W., & Jerjen, H. 2007, MNRAS, 378, 1575

Sérsic, J.-L. 1963, Bol Asoc Argent Astron, 6, 41

Sérsic, J. L. 1968, Atlas de Galaxias Australes (Córdoba: Argentina Observatorio Astronomico)

Sellwood J. A., & Wilkinson, A. 1993, Rep. Prof. Phys., 56, 173

Seth, A. C., Blum, R. D., Bastian, N., Caldwell, N., & Debattista, V. P. 2008, ApJ, 687, 997

Shapley, H. 1928, Nature, 122, 482

Shapley, H., & Swope, H. H. 1924, in Studies of the Galactic Center (Harvard: Astronomical Observatory of Harvard College), reprint 51 and 52

Shaviv G. 2011, in The Quest for Chemical Element Genesis and What the Chemical Elements Tell about the Universe, Springer Pub. Heidelberg (arXiv:1107.0442)

Shaw, M. A., & Gilmore, G. 1989, MNRAS, 237, 903

Shields, J. C., Walcher, C. J., Böker, T., Ho, L. C., Rix, H.-W., & van der Marel, R. P. 2008, ApJ, 682, 104

Simien, F., & de Vaucouleurs, G. 1986, ApJ, 302, 564

Simonneau, E., & Prada, F. 2004, Revista Mexicana de Astronomia y Astrofisica 40, 69 (astro-ph/9906151)

Skelton, R. E., Bell, E. F., & Somerville, R. S. 2009, ApJ, 699, L9

Slipher, V. M. 1917, Proc Am Philos Soc, 56, 403

Smith, S. 1935, ApJ, 82, 192

Smith Castelli, A. V., Faifer, F. R., Richtler, T., & Bassino, L. P. 2008, MNRAS, 391, 685

Soifer, B. T., & Neugebauer, G. 1991, AJ, 101, 354

Somerville, R. S., Gilmore, R. C., Primack, J. R., & Dominguez, A. 2011, MNRAS, submitted (arXiv:1104.0669)

Sparks, W. B. 1988, AJ, 95, 1569

Spano, M., Marcelin, M., Amram, P., Carignan, C., Epinat, B., Hernandez, O. 2008, MNRAS, 383, 297

Spergel, D. N. 2010, ApJS, 191, 58

Spinrad, H., et al. 1978, ApJ, 225, 56

Stebbins, J., & Whitford, A. E. 1934, PNAS, 20, 93

Steinmetz, M., & Navarro, J. F. 2002, New Astron, 7, 155

Stewart, K. R., Bullock, J. S., Wechsler, R. H., Maller, A. H., & Zentner, A. R. 2008, ApJ, 683, 597

Stiavelli, M., Miller, B. W., Ferguson, H. C., Mack, J., Whitmore, B. C., & Lotz, J. M. 2001, AJ, 121, 1385

Strom, K. M., & Strom, S. E. 1978, AJ, 83, 1293

Tal, T., & van Dokkum, P. G. 2011, ApJ, 731, 89

Temi, P., Brighenti, F., Mathews, W. G., & Bregman, J. D. 2004, ApJS, 151, 237

Terzić, B., & Graham, A. W. 2005, MNRAS, 362, 197

Terzić, B., & Sprague, B. J. 2007, MNRAS, 377, 855

Thakur, P., Ann, H. B., & Jiang, I.-G. 2009, ApJ, 693, 586

Thomas, D., & Davies, R. L., 2006, MNRAS, 366, 510

Tollerud, E. J., Bullock, J. S., Graves, G. J., & Wolf, J. 2011, ApJ, 726, 108

Tolstoy, E., Hill, V., & Tosi, M. 2009, ARA&A, 47, 371

Tonini, C., Maraston, C., Thomas, D., Devriendt, J., & Silk, J. 2010, MNRAS, 403, 1749

Tonry, J. 1981, ApJ, 251, L1

Tran, H. D., Tsvetanov, Z., Ford, H. C., Davies, J., Jaffe, W., van den Bosch, F. C., & Rest, A. 2001, AJ, 121, 2928

Tremonti, C. A., et al. 2004, ApJ, 613, 898

Trimble, V. 1999, Bull Am Astron Soc, 31, 1479 (#74.09)

Trujillo, I., Graham, A. W., & Caon, N. 2001, MNRAS, 326, 869

Trujillo, I., Asensio Ramos, A., Rubiño-Martín, J. A., Graham, A. W., Aguerri, J. A. L., Cepa, J., & Gutiérrez, C. M. 2002, MNRAS, 333, 510

Trujillo, I., Erwin, P., Asensio Ramos, A., & Graham, A. W. 2004, AJ, 127, 1917

Trujillo, I., et al. 2006, MNRAS, 373, L36

Trumpler, R. J. 1930a, PASP, 42, 214

Trumpler, R. J. 1930b, Lick Obs Bull, 14, 154

Tsikoudi, V. 1979, ApJ, 234, 842

Tully, R. B., & Fisher, J. R. 1977, A&A, 54, 661

Tully, R. B., Pierce, M. J., Huang, J.-S., Saunders, W., Verheijen, M. A. W., & Witchalls, P. L. 1998, AJ, 115, 2264

Tyson, J. A. 2002, Proc SPIE, 4836, 10

Unterborn, C. T., & Ryden, B. S. 2008, ApJ, 687, 976

van Albada, T. S. 1982, MNRAS, 201, 939

van den Bergh, S. 1977, in Evolution of Galaxies and StellarPopulations,ed.B.M.Tinsley,&R.B.Larson (New Haven: Yale University Observatory), 19

van den Bergh, S. 1986, AJ, 91, 271

van den Bergh, S. 1991, PASP, 103, 609

van den Bergh, S. 2008, A&A, 490, 97

van der Kruit, P. C. 1979, A&AS, 38, 15

van der Kruit, P. C. 1987, A&A, 173, 59

van der Kruit, P. C. 1988, A&A, 192, 117

van der Kruit, P. C., & Searle, L. 1981, A&A, 95, 105

van Houten, C. J. 1961, Bull Astron Inst Neth, 16, 1

Vennik, J., & Richter, G. M. 1994, Astron Nachr, 315, H3, 245

Vennik, J., Hopp, U., Kovachev, B., Kuhn, B., & Elsässer, H. 1996, A&ASS, 117, 261

Volonteri, M., Madau, P., & Haardt, F. 2003, ApJ, 593, 661

Wadadekar, Y., Robbason, B., & Kembhavi, A. 1999, AJ, 117, 1219

Wainscoat, R. J., Freeman, K. C., & Hyland, A. R. 1989, ApJ, 337, 163

Walcher, C. J., et al. 2005, ApJ, 618, 237

Walmsley, C. M., Bertout, C., Combes, F., Ferrara, A., Forveille, T., Guillot, T., Jones, A., & Shore, S. 2010, A&A, 518, 1

Webb, C. J. 1964, AJ, 69, 442

Weiner, B. J., Williams, T. B., van Gorkom, J. H., & Sellwood, J. A. 2001, ApJ, 546, 916

Weinzirl, T., Jogee, S., Khochfar, S., Burkert, A., & Kormendy, J. 2009, ApJ, 696, 411

Werner, M. W., et al. 2004, ApJS, 154, 1

White, R. E., III, Keel, W. C., & Conselice, C. J. 2000, ApJ, 542, 761

Williams, M. J., et al. 2010, MNRAS, 414, 2163

Wilson, C. P. 1975, AJ, 80, 175

Wirth, A., & Gallagher, J. S. 1984, ApJ, 282, 85

Wolf, M. 1908, Publ Astrophys Inst Kñig Heidelb, 3, 3

Wolf, J., Martinez, G. D., Bullock, J. S., Kaplinghat, M., Geha, M., Muñoz, R. R., Simon, J. D., & Avedo, F. F. 2010, MNRAS, 406, 1220

Wolfe, A. M., & Burbidge, G. R. 1970, ApJ, 161, 419

Wozniak, H., Friedli, D., Martinet, L., Martin, P., & Bratschi, P. 1995, A&AS, 111, 115

Wright, T. 1750, in An Original Theory or New Hypothesis of the Universe, (London: Macdonald)

Wyse, R. F. G., & Gilmore, G. 1995, AJ, 110, 2771

Wyse, R. F. G., Gilmore, G., & Franx, M. 1997, ARA&A, 35, 637

Yip, C.-W., Szalay, A. S., Carliles, S., & Budavari, T. 2011, ApJ, in press (arXiv:1101.5651)

York, D. G., et al. 2000, AJ, 120, 1579

Yoshizawa, M., & Wakamatsu, K. 1975, A&A, 44, 363

Young, C. K., & Currie, M. J. 1994, MNRAS, 268, L11

Young, C. K., & Currie, M. J. 1995, MNRAS, 273, 1141

Young, P. J., Westphal, J. A., Kristian, J., Wilson, C. P., & Landauer, F. P. 1978, ApJ, 221, 721

Zhao, H. 1996, MNRAS, 278, 488

Zibetti, S., White, S. D. M., Schneider, D. P., & Brinkmann, J. 2005, MNRAS, 358, 949

Zinnecker, H., Keable, C. J., Dunlop, J. S., Cannon, R. D., & Griffiths, W. K. 1988, in IAU Symp. 126, Globular Cluster Systems in Galaxies, ed. J. E. Grindlay, & A. G. D. Philip (Dordrecht: Kluwer), 603

Zoccali, M. et al. 2006, A&A, 457, L1

3 Star Formation in Galaxies

Samuel Boissier
Laboratoire d'Astrophysique de Marseille, Université Aix-Marseille
& CNRS, UMR7326, Marseille cedex 13, France

1 Introduction . . . 143

2 Theoretical Background . . . 145
2.1 Formalism . . . 145
2.2 Conditions for Star Formation . . . 145
2.2.1 Gravitational Instability . . . 146
2.2.2 Shear Criterion . . . 147
2.2.3 Formation of a Cold Phase . . . 147
2.3 Galactic Influences on the Star Formation Rates . . . 147
2.3.1 Free Fall . . . 148
2.3.2 Hydrostatic Equilibrium . . . 148
2.3.3 Gravitation Versus Dispersion . . . 148
2.3.4 Self-regulated Star Formation . . . 149
2.3.5 Cloud Collapse Versus Stellar Disruption . . . 149
2.3.6 Cloud–Cloud Collisions . . . 149
2.3.7 Physics of the ISM . . . 150
2.3.8 Influence of the Spiral Arms? . . . 151
2.3.9 Galactic Influences on the SFR: A Tentative Summary . . . 151
2.4 Starbursts and Peculiar Star Formation Regimes . . . 152

3 Measuring Star Formation Rates . . . 153
3.1 Proto-Stars, Young Stars and Stellar Remnants . . . 153
3.2 Stellar Continuum . . . 154
3.3 Recombination Emission Lines . . . 155
3.3.1 Motivation and General Approach . . . 155
3.3.2 Application to Hα . . . 157
3.3.3 Lyman α . . . 157
3.4 The Role of Dust . . . 158
3.4.1 Extinction Corrections from Recombination Lines . . . 158
3.4.2 Extinction Corrections in the UV . . . 158
3.4.3 Other Considerations on the Dust Attenuation . . . 159
3.4.4 Far-Infrared Star Formation Rates . . . 159
3.4.5 The Mixed Tracers . . . 161
3.5 Other Spectral Diagnostics . . . 161
3.5.1 [OII] 3,727 Å Forbidden Line . . . 161
3.5.2 [CII] 158 μm Fine structure line . . . 162
3.6 Radio Emission . . . 162

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_3, © Springer Science+Business Media Dordrecht 2013

3.7 X-Ray Luminosity 163
3.8 Additional Factors 164
3.8.1 The Effect of the Metallicity 164
3.8.2 Choice of the IMF 164
3.8.3 Effect of the Star Formation Micro-History 165

4 Star Formation Observed in Galaxies 166
4.1 Star Formation in the Local Galaxies 166
4.2 The Schmidt Laws 166
4.2.1 Preliminary Considerations: Many "Laws" 166
4.2.2 Which Law Is Right? 170
4.2.3 Which Scale Is Right? 171
4.2.4 Schmidt Laws: Current Observational Status 171
4.2.5 Starbursts and the Schmidt Law 174
4.3 Observed Thresholds 175
4.4 Relations to the Stellar Content: The Specific Star Formation Rate 176
4.5 Star Formation History 176

Acknowledgments 179

References 179

Abstract: The process of star formation is at the core of the evolutionary cycle of galaxies, as newborn stars produce new chemical elements, dust, and light. The energetic output delivered first by stellar winds and then by supernovae a few Myr after a star formation episode may also directly impact on the evolution of galaxies and their interstellar medium (ISM), as well as having an effect on the intergalactic medium (IGM), through feedback and outflows.

This chapter concerns star formation on galactic scales. First, the galactic processes that may affect large-scale star formation are presented. Second, the various methods to measure star formation rates are discussed (star formation tracers, timescales, calibrations, limits). Finally, the observational status concerning star formation in galaxies (its relation to other quantities and its evolution) is presented. The Schmidt Law (star formation rate–gas relationship) is amply discussed.

Keywords: Dust extinction, Evolution, Galaxy, Initial Mass Function, Schmidt Law, Star formation, Star formation diagnostics, Star formation history, Star formation rate, Star formation threshold, Star formation tracers, Toomre parameter

List of Abbreviations: *FIR*, Far infrared; *GMC*, Giant molecular clouds; *IGM*, Intergalactic medium; *IMF*, Initial mass function; *ISM*, Interstellar medium; *SFH*, Star formation history; *SFR*, Star formation rate; *SSFR*, Specific star formation rate; *YSO*, Young stellar objects

1 Introduction

Star formation is at the core of the cycle of evolution of galaxies. From their gas reservoir (and its replenishment), stars are formed at a given rate (the star formation rate, SFR) with a paramount impact on many aspects of galaxy evolution. The first one is the simple fact that galaxies are filled with stars that emit the light that allows us to see them. Generations after generations, the most massive of newly formed stars will also quickly yield the chemical elements formed in their core (stellar nucleosynthesis) to enrich the interstellar medium (ISM). This chemical enrichment will affect the properties and evolution of the next generations of stars and allow in fact the galaxies to hosts planets and life itself (both relying on heavy elements). Star formation is important for other aspects of extragalactic physics: the energy released by evolved stars and supernovae can affect the interstellar medium. It may heat the gas and prevent the collapse of the gas or, on the contrary, shock it and induce the formation of new stars. Its role on the distribution of stars and gas impacts the morphology of galaxies. Star formation finally affects the intergalactic medium (IGM), as winds due to violent episodes of star formation may exit their original galaxy and chemically and dynamically impact the IGM.

Because star formation is so crucial in shaping the galaxies (and maybe even their surroundings) the way they are seen, the subject has naturally motivated a lot of work. While it is generally admitted that stars form within molecular clouds, in the densest phases of the interstellar gas, the complex details of the process are not yet fully understood. For the actual physics of star formation, on small scales, the reader is referred to ➲ Chap. 3 of this handbook, or the McKee and Ostriker (2007) review. Such studies of the details of the process are accessible mostly in the Milky Way where star formation regions can be studied in great details.

On the other hand, star formation is also related to the scales of galaxies themselves. First, from a purely observational point of view, in many cases it is only possible to know the integrated SFR for the whole galaxy. But there are also physical links to be investigated: what is the main driver of star formation on the large scales? How is the SFR related to the galaxy gas reservoir? Is the SFR affected (enhanced or reduced) by the spiral arms of disk galaxies? By the density? By the galaxy kinematics? By other phenomenon? These questions are the subject of this chapter, focused on star formation on the scales of galaxies (by opposition to the scales of individual star-forming regions).

While early on the SFR (and its history) was derived in external galaxies from global quantities such as colors (with the help of evolutionary synthesis models), a number of SFR diagnostics (based on observables that will be called "SFR tracers") have been developed and largely used since the 1980s (emission lines, UV continuum, far-infrared luminosities). SFRs are usually deduced from these tracers using famous calibrations, relying on specific physics or empirical relations that are often forgotten. None of the SFR tracers in fact provide a really instantaneous SFR. To measure it, it is assumed that the SFR was constant on a typical timescale (between a few Myr and a few 100 Myr depending on the tracer). A given initial mass function (IMF) has also to be chosen. It is important to understand the origin of the calibrations used to derive the SFR and to know the assumptions underlying them and their limits.

Early on, Schmidt (1959) found a relation between the SFR and the gas density in the Milky Way. This relation, now known as the Schmidt Law, has been revisited frequently and has taken many forms (over the scales of whole galaxies or "local" scales of a few 100 pc within galaxies). It is still today an active field of research and a constraint for models of the evolution of galaxies. Today, SFR are routinely derived for large samples of galaxies, in the nearby universe, as well as in deep fields showing us galaxies in a young universe. The SFR distribution is computed and even integrated to provide a "cosmic" SFR and its evolution: the history of star formation in the whole universe. These studies bring an unprecedented amount of information on the history of galaxies.

In this chapter, the subjects briefly described above are distributed in three sections. In ❯ Sect. 2, the formalism describing star formation is presented, followed by the theoretical ideas suggesting relations between star formation and other galactic phenomenon or quantities. The diagnostics for measuring the star formation activity in galaxies (the SFR tracers) are presented in ❯ Sect. 3 including the underlying assumptions, the calibrations, the timescales, and some observational difficulties. Finally, a review of the current observational status concerning star formation on the scales of galaxies, the Schmidt Law, the specific star formation rate, and the star formation history is proposed in ❯ Sect. 4.

Other reviews (complementary to the material presented in these pages) can be found in books and classical papers, such as Chapter 9 of Mo et al. (2009), Kennicutt (1998a), and Larson (1992). A number of ideas about star formation in galaxies were discussed during a nice international conference hold at the Abbazia di Spineto on the subject "SFR@50: Filling the Cosmos with Stars" at the occasion of the 50th birthday of the seminal paper by Schmidt (1959). While no traditional proceedings were published, the reader may find the presentations given in that occasion on DVD or via the Internet.[1]

[1] http://www.arcetri.astro.it/sfr50/index.html

2 Theoretical Background

2.1 Formalism

In order to describe star formation in quantitative terms, it is necessary to adopt a simple formalism. It would be ideal to know at any time the functional form form(m, t) describing the number of stars dN formed in the dm mass interval during the time dt so that $dN =$ form$(m, t)dtdm$. It is convenient to decompose this functional form into two independent terms. First, the total amount of material forming stars from interstellar gas by unit of time (in, e.g., $M_\odot$ year^{-1}) is called the star formation rate (SFR, noted ψ in the following). Second, the stellar mass (m) distribution of a generation of newly formed stars (the initial mass function, IMF, noted ϕ in the following). The IMF is defined as $\phi(m) = dN/dm$ and normalized as follows:

$$\int_{M_l}^{M_u} m\phi(m)dm = 1, \tag{3.1}$$

where M_l and M_u are the lower and upper mass limits, respectively. In his classical study, Salpeter (1955) defined in fact the logarithmic IMF $\xi(m) = dN/d\log(m)$. The two functions are simply linked by $\xi(m) = ln(10)m\phi(m)$. There are some uncertainties concerning the upper and lower limits, but it is usual to consider values between 0.1 and 100 $M_\odot$. With this formalism, form$(m, t) = \psi(t)\phi(m)$, where it is implicitly assumed that the IMF is stationary in time and only dependent on m (see however, ➜ Sect. 3.8.2 for different suggestions).

The IMF is often parametrized as a power law, following the classical study of Salpeter (1955). Under this assumption, it is possible to write $\phi(m) \propto m^{-(1+x)}$ or equivalently $\xi(m) \propto m^{-x}$. The index x, measuring the logarithmic slope of the IMF, was found by Salpeter (1955) to have a value of 1.35 in the solar neighborhood. Even though it was estimated over a limited mass range, this value is still frequently used. However, more recent determination of the IMF have established that the IMF presents a flattening at low masses. The universal IMF of Kroupa (2001) is an example of formulations more in accordance with the recent measurements. It is characterized by a slope of $x = 0.3$ between 0.1 and 0.5 $M_\odot$, and $x = 1.3$ for masses larger than 0.5 $M_\odot$. In the following, this IMF (largely used in the literature) is adopted. More details on the IMF can be found elsewhere in these volumes (➜ Chap. 4 of Volume 5).

The IMF might result from fundamental physical processes (such as turbulence and self-gravity, etc.) leading to star formation. For this reason, it is reasonable to expect some level of universality, and there are indeed good signs of it (e.g., Bastian et al. 2010, and references therein), but see other considerations in ➜ Sect. 3.8.2. With a universal IMF, star formation is summarized in the SFR, which dictates how much mass is converted into stars per unit time and, as a consequence, what will be the production of heavy elements, light, feedback, etc. Understanding what influences the SFR on the scales of galaxies is thus of paramount importance for every aspect of galaxy evolution. The rest of the chapter is an attempt to summarize the present understanding of the SFR on the galactic scales.

2.2 Conditions for Star Formation

The modern view of star formation is that giant molecular clouds (GMC) are a prerequisite to form stars (e.g., Leroy et al. 2008, and references therein). Once molecular clouds are formed, "local" processes (not considered in this chapter) will transform part of them into stars, on a short timescales (e.g., Tamburro et al. 2008). In that case, the question of star formation reduces to the formation of molecular gas and GMCs from which stars will be formed (e.g., Blitz and Rosolowsky 2006).

Many processes reviewed below have been presented as favoring the collapse or condensation of gas. In the current adopted view, they should be seen as not leading directly to the apparition of stars, but to the formation of molecular gas and molecular clouds. The processes described in the ➋ Sect. 2.2 concern conditions that should be fulfilled for star formation to occur, leading to the concept of thresholds for star formation. ➋ Section 2.3 will concern the direct influences on the SFR itself.

2.2.1 Gravitational Instability

The most obvious and discussed effect is the disk gravitational instability. In reality, other instabilities (e.g., the thermal and Parker instability, see for instance in Elmegreen 1993a) may play a role. Wang and Silk (1994) argue that the gravitational one is the leading phenomenon, with only some contribution from other instabilities, an idea frequently accepted. While studying the stability of stellar disks (in order to explain the observation of spiral and S0 galaxies with smooth distribution of stars), Toomre (1964) established that a rotating stellar disk is unstable when $Q_* < 1$, where the "Toomre parameter" Q_* is defined as

$$Q_* = \frac{\sigma_* \kappa}{3.36 G \Sigma_*}. \tag{3.2}$$

Σ_* is the stellar surface density, σ_* is the velocity dispersion, and κ is the epicyclic frequency. κ can be defined as (Binney and Tremaine 1987):

$$\kappa = \left(R \frac{d\Omega^2}{dR} + 4\Omega^2 \right)^{0.5}. \tag{3.3}$$

The Toomre parameter Q_* can be seen as the balance between the velocity dispersion and the Coriolis force tending to tear apart the stars, counterbalancing their gravity. This result inspired a lot of other works that gave to this "Toomre stability criterion" a diversity of forms. For a gaseous disk (of surface density Σ_{gas} and velocity dispersion σ_{gas}), one can assume (e.g., Cowie 1981; Kennicutt 1989; Wang and Silk 1994) that the disk is unstable when $Q < 1$ where

$$Q = \frac{\sigma_{\text{gas}} \kappa}{\pi G \Sigma_{\text{gas}}}. \tag{3.4}$$

It is common to define a gas "critical" density Σ_{crit} such that $Q = 1$. The gaseous disk is unstable when $\Sigma_{\text{gas}} > \Sigma_{\text{crit}}$ following the idea of Quirk (1972). While originally, the criterion concerns axisymetric instabilities, it may also be appropriate for various processes of star formation, for instance, the collapse of expanding shells or of turbulence-compressed regions (see Elmegreen and Hunter 2006, and references therein).

In a number of studies (e.g., Kennicutt 1989; Martin and Kennicutt 2001), a pragmatic approach has been adopted by using Q' in place of Q, defined simply as

$$Q' = \alpha_Q Q, \tag{3.5}$$

α_Q is a normalization factor, determined by measuring the limit of the disk stability and affecting $Q' = 1$ to it. In this way, Martin and Kennicutt (2001) found $\alpha_Q = 0.69 \pm 0.2$.

Wang and Silk (1994) also proposed an approximation for the Q parameter taking into account the effect of the stellar disk (based on the analysis of two isothermal fluids in Jog and Solomon 1984) whose density can help the collapse of the gaseous component:

$$Q \simeq \frac{\sigma_{\rm gas}\kappa}{\pi G \Sigma_{\rm gas}} \left(1 + \frac{\Sigma_* \, \sigma_{\rm gas}}{\Sigma_{\rm gas} \, \sigma_*}\right)^{-1} . \tag{3.6}$$

Note that this effect of the stellar density in (❯ 3.6) can be hidden in α_Q of (❯ 3.5) that should then however depend on the stellar density.

2.2.2 Shear Criterion

Noticing that star formation seems to occur in regions of low shear, Elmegreen (1993a) proposed another definition of the Q parameter, incorporating the shear rather than the epicyclic frequency:

$$Q_A = \frac{2.5\sigma_{\rm gas} A}{\pi G \Sigma_{\rm gas}}, \tag{3.7}$$

where A is the Oort shear constant:

$$A = 0.5 R \frac{d\Omega}{dR}. \tag{3.8}$$

Ω is the angular frequency. For a flat rotation curve with a constant linear velocity V (independent from the radius R), one has $\Omega = V/R$. It is easy to find that Q from (❯ 3.4) is then close to Q_A. This is true for a large part of disks in spiral galaxies, but makes differences in the inner parts of spirals and dwarf galaxies where the shear is low (Hunter et al. 1998). Seigar (2005) argued that the observed shear rates do present a threshold (at $A/\Omega \sim 0.7$) corresponding to a null star formation rate.

2.2.3 Formation of a Cold Phase

Schaye (2004) suggested that the formation of a cold phase and an efficient formation of molecular gas (depending on the physics of the ISM) is more relevant to star formation than gravitational stability effects over large scale. He found a critical surface density that depends (weakly) on the metallicity, the gas fraction, the flux of ionizing photons, and the ratio of the thermal to total pressure. He provided a fitting formula in his paper, and a typical local critical gas density $\Sigma_{\rm crit} \sim 3 - 10 {\rm M}_\odot\,{\rm pc}^{-2}$ for "reasonable values" of the parameters.

2.3 Galactic Influences on the Star Formation Rates

Many processes on galactic scales have been proposed to directly affect the star formation rate itself. The more often discussed ones are reviewed below. A comparative study of many propositions has been done recently by Leroy et al. (2008).

2.3.1 Free Fall

It is possible to write the star formation rate density in a disk as the ratio of the gas density and the timescale to turn gas into stars, with an efficiency ϵ (e.g., Wang and Silk 1994; Larson 1992):

$$\Sigma_\psi = \epsilon \frac{\Sigma_{\text{gas}}}{\tau}. \tag{3.9}$$

What sets the timescale τ? The first natural idea is to consider the free fall time for pure gravitational instability that leads to $\tau \propto \rho_{\text{gas}}^{-0.5}$ (e.g., Madore 1977), giving for a constant scale height of the disk (e.g., Bigiel et al. 2008; Leroy et al. 2008, and references therein):

$$\Sigma_\psi \propto \Sigma_{\text{gas}}^{1.5}. \tag{3.10}$$

2.3.2 Hydrostatic Equilibrium

Contrary to the assumption in the above section, the scale height (h) of a disk might not be constant, but regulated by the hydrostatic equilibrium of the disk. It is then approximated by Corbelli (2003) as:

$$h = \frac{\sigma_{\text{gas}}}{\pi G} \left(\frac{\Sigma_{\text{gas}}}{\sigma_{\text{gas}}} + \frac{\Sigma_*}{\sigma_*} \right)^{-1}. \tag{3.11}$$

It is still possible to use $\tau \propto \rho_{\text{gas}}^{-0.5}$ with $\rho_{\text{gas}} = \Sigma_{\text{gas}}/2h$, leading to (see also Leroy et al. 2008):

$$\Sigma_\psi \propto \frac{\Sigma_{\text{gas}}^2}{\sigma_{\text{gas}}} \left(1 + \frac{\Sigma_*}{\Sigma_{\text{gas}}} \frac{\sigma_{\text{gas}}}{\sigma_{*,z}} \right)^{0.5} \tag{3.12}$$

In a parallel approach, Abramova and Zasov (2008) used the assumption of hydrostatic equilibrium to compute the scale height of the gaseous disk, and then use it to compute volume density. They showed that the volume density of gas and SFR correlate better than the surface densities. They also suggested a direct dependence of the SFR on the stellar surface density.

2.3.3 Gravitation Versus Dispersion

Another point of view is to consider that the perturbation growth timescale τ from (➲ 3.9) is obtained by balancing dispersion against gravitation (Larson 1992):

$$\tau \propto \frac{\sigma_{\text{gas}}}{\pi G \Sigma_{\text{gas}}}. \tag{3.13}$$

Assuming the velocity dispersion is constant, the star formation rate density is then

$$\Sigma_\psi \propto \Sigma_{\text{gas}}^2. \tag{3.14}$$

This relation should apply to kpc scales because the gravitational aggregation proceeds on these scales.

2.3.4 Self-regulated Star Formation

When star formation occurs, it may dynamically heat the gaseous disk. Then, σ_{gas} in (3.13) is not constant (in contrast with the assumption of Sect. 2.3.3) but will increase with star formation, what will also increase the Toomre parameter Q. The disk may then become stable, star formation stops. The gas can then cool down and Q go down. Thus, it is possible to consider that Q has a quasi-constant value (around 1). Self-regulation was suggested by, e.g., Kennicutt (1989) noticing that the ratio $\Sigma_{\text{gas}}/\Sigma_{\text{crit}}$ presents a small variation over a large range of Σ_{gas}. In that case, by combining $Q = 1$, (3.4) and (3.13), one obtains $\tau \propto \kappa^{-1}$, leading for the star formation rate (using (3.9)) to

$$\Sigma_{\psi} \propto \Sigma_{\text{gas}}\kappa. \tag{3.15}$$

Note that for a flat rotation curve (a good approximation to observations of rotation curves in the Milky Way and nearby galaxies), $\kappa = \sqrt{2}\Omega$ so that the SFR depends in this case on the density of gas and the rotation curve, both quantities being relatively easy to observe in nearby galaxies.

Larson (1992) noted that the cases leading to (3.15) and (3.14) are two simplistic assumptions and that reality might be somewhere in between, i.e.,

$$\Sigma_{\psi} \propto \Sigma_{\text{gas}}^{\alpha}\kappa^{\beta}; 1 < \alpha < 2; 0 < \beta < 1. \tag{3.16}$$

For a flat rotation curve, this can be written as

$$\Sigma_{\psi} \propto \Sigma_{\text{gas}}^{\alpha}\Omega^{\beta}. \tag{3.17}$$

Note that this self-regulated form is obtained for either a constant Toomre Q (3.4) or for a constant value of the shear Toomre parameter Q_A (3.7).

2.3.5 Cloud Collapse Versus Stellar Disruption

Madore (2010) proposed that the collapse time scale for a cloud (parametrized as $\tau_C \propto \rho_{\text{gas}}^{-n}$) should be combined with a timescale τ_S, characteristic of the disruptive effect of star formation (at a place in a galaxy, once stars are formed, the gas is dispersed and ionized, so that no further star formation can occur at that place for the time τ_S). The star formation rate (per volume unit) can then be written as

$$\rho_{\psi} \propto \frac{\rho_{\text{gas}}^{n}}{\tau_S + \rho_{\text{gas}}^{-n}}. \tag{3.18}$$

Madore (2010) showed that this functional form allow to reproduce the trend obtained by Bigiel et al. (2008) between the star formation rate and the total gas surface density (see Sect. 4).

2.3.6 Cloud–Cloud Collisions

Assuming that the crucial step of neutral gas turning into molecular gas (formation of GMCs) occur during cloud-cloud collisions, Wyse (1986) proposed that the star formation rate should be proportional to Σ_{HI}^{2} and included a dynamical factor to take into account the effect of the spiral arms (see Sect. 2.3.8 below).

Under the assumption of cloud-cloud collisions, Tan (2000) obtained a more complex formula, including the effect of shear on the collision rate:

$$\Sigma_{\psi} \propto \Sigma_{\rm gas}\Omega(1 - 0.7\beta), \tag{3.19}$$

where $\beta = d\,ln(V)/d\,ln(R)$. Note that β is null for a flat rotation curve, in wich case the SFR law obtained has the same form as (➋ 3.17) obtained with a different approach.

2.3.7 Physics of the ISM

Since stars are formed within molecular clouds, the SFR may be directly linked to the processes affecting the amount of molecular gas present in a galaxy, i.e., to the detailed physics of the ISM. Which process actually governs the molecular fraction within galaxies is however under discussions.

Elmegreen (1993b) proposed that the fraction of molecular gas is set by the hydrostatic pressure and the radiation field and that this is the limiting factor for star formation rather than the gravitational assembly of material into clouds and GMCs. This idea was pursued by several authors (Leroy et al. 2008, and references therein). Blitz and Rosolowsky (2006) expressed it by saying that the molecular ratio $R_{\rm mol} = \Sigma_{H2}/\Sigma_{HI}$ should depend on the pressure:

$$R_{\rm mol} = (P/P0)^{\alpha}. \tag{3.20}$$

For low pressures ($P \ll P0$), over large part of galaxies (where HI dominates over H2), the SFR should then follow a relation of the type:

$$\Sigma_{\rm SFR} \propto \Sigma_{\rm gas}(P/P0)^{\alpha}. \tag{3.21}$$

Blitz and Rosolowsky (2006) found that their observations in 14 galaxies (including the Milky Way) were in good agreement with such a relation with an index close to $\alpha \sim 0.9$, also found in other studies (see Leroy et al. 2008, and references therein). Blitz and Rosolowsky (2006) noticed however that under their assumptions (constant velocity gas dispersion, turbulent pressure providing support), their expression is equivalent to $R_{\rm mol}$ being a function of the midplane gas density and that their data alone cannot distinguish these possibilities. They stressed that while they suggest that the hydrostatic pressure determines the molecular fraction, it does not determine how the GMC are assembled from molecular gas, or if the gas is first assembled into GMC-sized object, and then becomes molecular. In other words, even if they argue that the hydrostatic pressure plays a major role, there is room for other processes in the formation of GMCs and then, likely, of stars.

Monaco et al. (2012) found that their multiphases models including a pressure-determined molecular fraction do produce a dynamical relation ($\Sigma_{\rm SFR} \propto \Sigma_{\rm gas}\kappa$) that they interpret as resulting from the equilibrium between the energy injection (via supernovae) and dissipation.

Krumholz et al. (2009) proposed an alternate model based on the combination of few simple ideas. The Hydrogen self-shielding determines the amount of molecular gas that once formed is regulated by internal process and form stars at a rate of 1% per free fall time (as a result of turbulence). This allows them to predict the $\Sigma_{\psi} - \Sigma_{\rm gas}$ relationship (as well as the molecular fraction) as a function of the metallicity of the gas.

2.3.8 Influence of the Spiral Arms?

The first impression when one looks at most nearby spiral galaxy is that young stars are distributed along spiral arms or arcs, and it is hard to think that star formation is unrelated to spiral structure. Elmegreen (1979) however suggested that two types of spiral arms should be considered. The spiral density wave, related to the stellar underlying density, is responsible for the grand design spirals that are often seen in galaxies. The other type is made of stochastic spiral arms or filaments produced by the shear of self-propagating star formation (this type of stochastic arms would be a product of star formation but would not affect the galaxy SFR). According to Elmegreen (1979), the density wave/grand design arms can play an active role by compressing or shocking the interstellar material. Cowie (1981) also found that the gaseous disk of the Milky Way is globally stable against collapse, but that the passage of the clouds ensemble trough the spiral arm will modify the velocity dispersion of the clouds and produce the instability. Star formation should then proceed only in the arms. Through such processes, grand-design spiral arms should affect star formation. If star formation (or cloud collapse) is indeed enhanced when the gas crosses the arms, then the efficiency of star formation should depend on the frequency at which it happens. On this basis, Wyse (1986) proposed that

$$\Sigma_{\psi} \propto (\Omega - \Omega_P). \tag{3.22}$$

Ω_P is the angular velocity of the spiral pattern. This term should in fact come in combination with other factors, depending on which of the processes adopted above is considered (e.g., one of the (➲ 3.10), (➲ 3.12), (➲ 3.14), (➲ 3.15), (➲ 3.19), and (➲ 3.21)). Considering that Ω_P of the density wave is small with respect to the angular frequency of the material, the factor above becomes simply Ω. Wyse (1986) combined it with a cloud-cloud assumption to compute the SFR as $\Sigma_{\psi} \propto \Sigma_{HI}^{2}(\Omega - \Omega_P)$, and it was generalized in Wyse and Silk (1989) in $\Sigma_{\psi} \propto \Sigma_{HI}^{n}\Omega$ (n = 1,2). Interestingly, in combination with the total gas density rather than HI, the spiral arm term (➲ 3.22) suggests again a form similar to (➲ 3.17) even if it is for a totally different reason! This form seems to be robust in the sense that it accommodates various theories.

Before closing this section, it should be noted that active star formation is present even when no strong spiral structure is present (e.g., Kennicutt 1989; Tan 2000), and no clear-cut differences in star formation rates between grand-design and flocculent spirals was ever established. Maybe after all, the differences induced by the grand design spiral arms are minor, organizing the star formation in a different way rather than enhancing it.

2.3.9 Galactic Influences on the SFR: A Tentative Summary

Many theoretical ideas lead to different relations between various quantities and the SFR. The problem is that some of these ideas actually drive to the same relationships (at least under some of the possible assumptions). For instance, the dynamical factor Ω can appear as a reference to spiral arms or instabilities of the disk. Also, the ISM phase balance can under some conditions be directly related to the gas density (and thus hide in another index of a dependence on Σ_{gas}). In the case of a self-regulated star formation rate (constant Q), the same expression can be obtained based on different Q definitions (collapse versus dispersion due to Coriolis forces or shear). Moreover, the proposed explanations are not totally exclusives: the self-regulation ($Q = 1$) could be view as an approximation of the H2/HI equilibrium regulated by SN feedback!

Following the approach of Larson (1992) already mentioned, it is possible to try to write a general form for the factors affecting the star formation rate, by combining all the previously described ones. For a flat rotation curve, this general form would be (willingly forgetting other possible factors, e.g., the stellar surface density)

$$\Sigma_\psi \propto \Sigma_{\text{gas}}^{\alpha} \Omega^{\beta} P^{\gamma} \tag{3.23}$$

(with α between 1 and 4, β between 0 and 2, γ between 0 and ~1). Unfortunately, these indexes are partly degenerated since, e.g., Σ_{gas} and Ω both decline with radius within galaxies. It will then be hard to derive them from observations alone.

Nevertheless, observational studies (see ➋ Sect. 4) will allow us to have a minimum of constraints on the possible relation between star formation and various quantities. This may help us to decide among the various processes at play which is the leading one (but it is already visible that this task will be hard). Even if the physical basis for these relations stay under debate, they will still provide useful constraints for models of evolution of galaxies.

2.4 Starbursts and Peculiar Star Formation Regimes

Up to this point, star formation has been discussed as a secular process occurring in a galactic disk. There are however different situations.

Starbursts regions (or galaxies) are characterized by elevated star formation rates, with an unusually high efficiency (e.g., Larson 1992, and references therein). Characterized by some signatures such as blue colors (Searle et al. 1973) and strong high-excitation emission lines, they form an ill-defined family of objects (e.g., Meurer et al. 1995, 1999) with large SFR (in absolute value, but also relative to the past average SFR, or in surface density) and short consumption timescales (the SFR is so high that the gas reservoir would be quickly depleted at this rate). As a result, the term apply to a number of objects that are not necessarily similar (especially when comparing high and low redshift ones).

Starbursts are often found in the nuclear regions of galaxies. This nuclear emission is found in an increasing fraction of galaxies with types going from early (8% in S0) to late (80% in Sc-Im) (Kennicutt 1998a, and references within). ➋ Eq. 3.15 suggests that the high level of star formation in central regions could be related to the elevated densities and high values of κ that are found there.

Starbursts are also revealed in the far infrared (FIR) with the observation of galaxies with very large luminosities (up to $10^{13} L_\odot$) corresponding to SFR up to 1,000 $\text{M}_\odot$ year^{-1}, occurring in very dense molecular gas, in which the optical radiation is almost totally absorbed. Kennicutt (1998a) reminds that the most luminous galaxies in FIR are systems in which a mass of gas corresponding to the total ISM of a normal galaxy is compressed into a small area and entirely transformed into stars.

It is tempting to relate the peculiar regime of starbursts to tidal encounters and mergers that could be responsible for enhancing the SFR by factor 10–100 (Kennicutt 1998a). Recent interactions and mergers simulations by, e.g., Di Matteo et al. (2008) produce a modest enhancement of the SFR (by a factor 5) but the modes of star formation during major mergers may need high-resolution simulations and hence be hard to catch in simulations in general (see, e.g., Teyssier et al. 2010, obtaining a gain of a factor 10 in star formation efficiency with high-resolution simulations). Barnes (2004) proposed that a shock-induced mode of star formation takes place during such interactions. This mode is usually not taken into account (and hard to model in general) but provided a good fit to the observations of the Mice system.

3 Measuring Star Formation Rates

Before discussing the various observational evidences and especially those linked to theoretical expectations of the previous section, it should be presented how the SFR is actually measured. A famous review concerning tracers of star formation can be found in Kennicutt (1998a) whose formula are widely used. The reader is referred to this paper for many interesting discussions on the various tracers of the SFR. Recently, Calzetti et al. (2009) reviewed the subject, focusing in the long wavelength range (infrared and radio).

Almost all SFR measurements rely on few principles and assumptions. The main idea is to have an observable linked to the amount of recent star formation, i.e., of stars formed within the last time interval δt, with δt being small with respect to the age of the galaxy. Since massive stars do have short lifetimes (that can be associated to δt), astronomers can use them (or their observable effects) to derive the SFR. Once the amount of massive stars formed during δt is known, it is easy to calculate the total amount of stars formed (in solar masses) by extrapolating over the full IMF and divide by δt to get a rate (the SFR).

Often, it is not possible to actually count massive stars. Instead, some emission (that will be called a "star formation tracer") due to their presence is measured, and the actual SFR is deduced from the flux detected through calibrations described below.

A purely "instantaneous" rate cannot be measured. All tracers have different typical timescales associated, depending on the mass range of the stars they are related to. Thus, the tracers are calibrated by assuming the SFR is constant over the typical timescale δt for this tracer (e.g., the lifetimes of stars responsible for it). In other words, an assumption of steady state is made to compute the SFR. Note that the SFR deduced; in this way is well an estimate assuming that the SFR was constant over the timescale δt, and not the average of the actual SFR during this period (➊ *Fig. 3-1* shows that the two are not equivalent).

Finally, in the following, "primary tracer" is used for measurements directly related to massive stars. The expression "secondary tracer" is used for those tracers that were introduced (at least historically) by empirically noting a correlation with a primary tracer that has been used to calibrate the secondary one. It is true that often more detailed models were constructed to recalibrate secondary tracers (at the price of more complex models), but these works were still motivated by the empirical evidence of a correlation in the first place.

3.1 Proto-Stars, Young Stars and Stellar Remnants

The Milky Way is a peculiar case in terms of SFR measurement. Most of the techniques detailed below cannot be applied, at least without some modifications. However, its SFR has been studied, mostly by using as SFR tracers the number of "young objects," as it was done by Schmidt (1959) early on. Guibert et al. (1978) provide a list of objects related to recent star formation that can be used to trace the SFR. It includes OB stars, giant HII regions, supernova remnants, and pulsars. In the recent years, the SFR in our galaxy was studied by detecting OB stars in

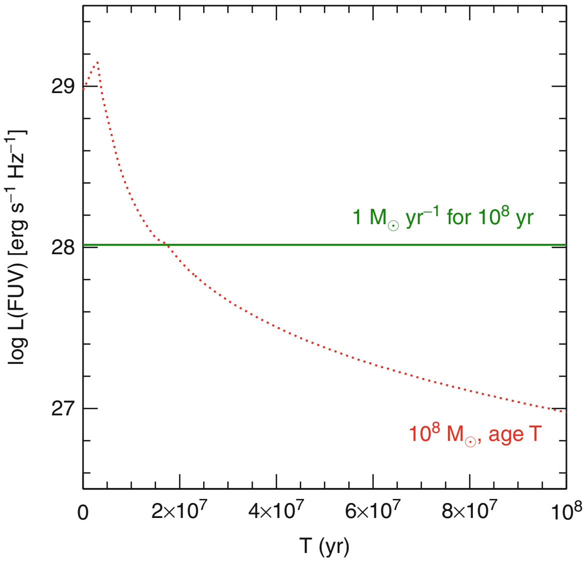


Fig. 3-1
The *dotted curve* shows the FUV luminosity (a SFR tracer) of a 10^8 M$_\odot$ instantaneous burst of age T as a function of T. The *horizontal line* shows the UV luminosity generated by a constant SFR of 1 M$_\odot$ year^{-1} during 10^8 year. All these populations have the same average SFR within 10^8 year, the timescale for the FUV emission. However, the SFR deduced from L(FUV) assuming a constant SFR can vary by 2 orders of magnitude

the far infrared. The ISM, being transparent at this wavelength, it is possible to detect star-forming regions as point-like source over the whole galaxy (e.g., Bronfman et al. 2000; Luna et al. 2006). It is also possible to detect young stellar objects (YSOs) and thus derive the SFR from pre-main-sequence objects themselves (see Robitaille and Whitney 2010).

Even if this "resolved approach" cannot be used in general for external galaxies, it is mentioned here because the spatial resolution of current and future observatories allow to contemplate the possibility to use a similar approach in our nearest neighbors (see, e.g., Whitney et al. 2008, for YSOs in the LMC).

3.2 Stellar Continuum

The continuum spectrum of a galaxy is the superposition of the emission of many generation of stars. In the UV, this emission is dominated by the hot young massive stars (above a few solar masses) and can thus be used to trace the "recent" star formation, over the lifetime of the contributing stars (of the order of 100 Myr).

For a steady state and a chosen IMF (and its mass limits), it is possible to relate the SFR and the luminosity of stars. Following Kennicutt (1998a), it can be written

$$\psi_\nu(\mathrm{M_\odot year^{-1}}) = C \times L_\nu(\mathrm{ergs^{-1}Hz^{-1}}). \qquad (3.24)$$

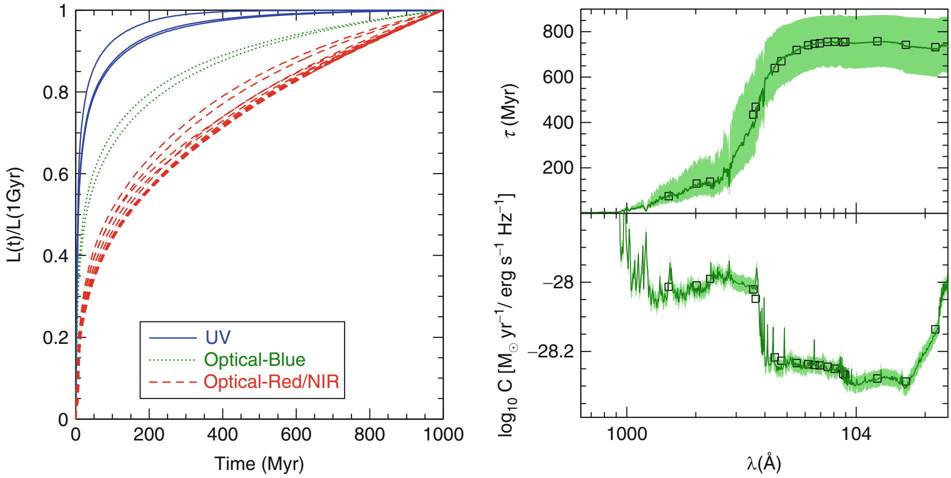


Fig. 3-2

***Left*: evolution of luminosity in a collection of photometric filters (in the UV: GALEX FUV, NUV, and FOCA; *optical/blue*: sloan u and Jonhson U; *optical/red/near infrared*: g, r, i, z, B, V, R, Rc, I, Ic, J, H, K) for a constant star formation rate of 1 $M_\odot$ year^{-1} for 10^9 years with the universal IMF of Kroupa (2001), computed using Starburst99. The luminosity is normalized by its final value. *Right-top*: time needed to reach 90% of the 1 Gyr luminosity as a function of the wavelength. (The *shaded area* shows how this curve shifts going from 85% to 95%.) *Right-bottom*: value of the calibration factor C at this time. The *squares* show the results for the same broadband filters as on the left**

The value of the conversion factor C can be computed with a spectral synthesis code, assuming a constant SFR. *Figure 3-2* shows (as a function of wavelength) the time needed for the luminosity to reach 90% of its 1 Gyr value for a constant SFR, as well as the value of *C* at that time. Because the SFR should be derived for massive stars with short lifetimes, this method should be used only in the UV, but the figure extends the results to the optical/near infrared for educative purpose. Short timescales are clearly obtained in the UV that makes a good tracer on ~100 Myr timescale. Fortunately, for reasonable IMFs, the L_ν luminosity per unit frequency in the UV has a relatively flat spectrum so that the calibration factor *C* depends weakly on the wavelength in this range. The various proposed calibrations however vary by 0.3 dex, depending on the stellar libraries used (Kennicutt 1998a). Calibration factors (and associated timescales) are provided in *Table 3-1* for several common photometric bands.

The "U" bands are intermediate: their flux to SFR conversion factor is close to the UV one, but the timescales are getting larger (about 500 Myr to the 90% level as defined previously) and thus are not as good to trace recent star formation.

3.3 Recombination Emission Lines

3.3.1 Motivation and General Approach

The strong flux emitted by young massive stars at shorter wavelengths than the Lyman limit ionizes the gas surrounding them. The recombination lines that follow can be used to estimate

Table 3-1

Primary SFR tracers. Timescales and coefficients are computed using Starburst99, solar metallicity, with the Kroupa (2001) "universal" IMF (0.1–100 M$_\odot$, slope x = 0.3/1.3 below/above 0.5 M$_\odot$). Values are given (adopting a constant SFR) when the luminosity reach 90% of the luminosity after 1 Gyr

Observation	Timescale	C	Comments
YSOs, stars, and remnants	Variable		Traditional method in the Milky Way, difficult in distant objects
		In (➜ 3.24)	
1,524 Å(FUV-GALEX)	75 Myr	0.97 10^{-28}	Easy at high redshift, but large/uncertain extinction
2,018 Å(UV-FOCA)	131 Myr	0.98 10^{-28}	
2,308 Å(NUV-GALEX)	139 Myr	1.02 10^{-28}	
3,561 Å(u)	434 Myr	0.95 10^{-28}	Longer timescales
3,651 Å(U)	469 Myr	0.89 10^{-28}	
		In (➜ 3.26)	
Hα	6.5 Myr	5.1 10^{-42}	Strong line, short timescale but difficulties with extinction, NII contamination, diffuse/absorbed fractions, sensitive to the upper IMF slope
Lyα	6.5 Myr		Lyα/Hα = 8.11 in theory but has to be multiplied by the unknown escape fraction

this ionizing flux and hence the SFR. The most famous SFR tracer among recombination lines is Hα (owing to its strength), but many lines from the Balmer, Pashen, and Brackett series have been used too.

For Hβ, it is possible to write (Osterbrock and Ferland 2006), assuming that all the photons are absorbed and reemitted:

$$L(H\beta) = h\nu_{H\beta} N_{\text{photons}} \frac{\alpha^{\text{eff}}_{H\beta}}{\alpha_B}, \tag{3.25}$$

where N_{photons} is the number of ionizing photons. N_{photons} can be obtained from stellar synthesis codes (such as Starburst99 for instance). For a temperature of 10,000 K and Osterbrock case B recombination, the recombination coefficients are $\alpha^{\text{eff}}_{H\beta} = 3.03\ 10^{-14}\text{cm}^3\text{s}^{-1}$, $\alpha_B = 2.59\ 10^{-13}\text{cm}^3\text{s}^{-1}$. Relations for other hydrogen recombination lines can be computed using the relative intensities of these lines with respect to $H\beta$ (Table 4.4 of Osterbrock and Ferland 2006). For instance, in the above-mentioned conditions, $L(H\alpha)/L(H\beta) = 2.847$, $L(P\alpha)/L(H\beta) = 0.332$, $L(P\beta)/L(H\beta) = 0.162$.

It is thus possible to use these relationships to compute the luminosity in any recombination line for a constant SFR and thus calibrate a relation between this luminosity and the SFR. For the same assumptions used for the stellar continuum (Kroupa 2001 IMF, constant SFR for 1 Gyr), the 90% level is reached in 6.5 Myr, almost instantaneously with respect to the long timescales obtained even with the UV continuum. This comes from the fact that the ionizing photons are on average emitted by more massive stars with shorter lifetimes than the one dominating the UV spectrum.

3.3.2 Application to Hα

Since Hα is a strong emission line with a short timescale, relatively easy to observe, moreover with good spatial resolution, it is an important tracer of star formation, and this section presents some more considerations about it. Using the method and formulae underlined in ➋ Sect. 3.3.1, it is possible to calibrate the relation between the Hα line intensity and the SFR:

$$\psi_{H\alpha}(\mathrm{M}_\odot\,\mathrm{year}^{-1}) = C \times L(H\alpha)(\mathrm{erg\,s}^{-1}). \tag{3.26}$$

The coefficient is given in ➋ *Table 3-1*, along with other SFR indicators.

Despite the advantages of Hα, making it the preferred SFR tracer, many uncertainties and difficulties exist. First, a dispersion between calibrations of 30% is due to various stellar models (e.g., Kennicutt 1998a), and the result is also very sensitive to the IMF slope in the very upper mass range. Hα, in order to be a precise tracer of the SFR must take into account the extinction that reduces the observed luminosity (see discussion on the role of dust in ➋ Sect. 3.4). When narrow-band filters around Hα are used to measure the luminosity of the line, it is also necessary to take into account the contamination of the Hα line by the neighboring [NII] λ 6548 and 6584 lines. In the absence of spectroscopy, typical ratios are used, see, e.g., the appendix of Boselli et al. (2001) for ratios given as a function of morphological type.

A fraction of the ionizing photons is in fact absorbed by dust rather than by the gas. Thus, the calibration should be corrected for this fraction that is not participating in the ionization. This fraction is estimated to be equal to 20% by Charlot et al. (2002), 0–30% by Bicker and Fritze-v. Alvensleben (2005) depending on the metallicity of the object.

Finally, a fraction of the ionizing photons may escape the HII regions. While in individual regions this escape fraction can be high (15–50%), it is thought to be lower on the scale of galaxies as a whole (Kennicutt 1998a, and references therein).

3.3.3 Lyman α

From the above, it is easy to find that the Lyα luminosity should be large, and thus should be a good tracer of star formation. This emission line has in fact been used in the recent years to search and find high redshift star-forming galaxies. The relations Lyα/Hβ = 23.1 and Lyα/Hα = 8.11 are obtained for a 10,000 K case B scenario (Osterbrock and Ferland 2006) and combined with, e.g., (➋ 3.26) should allow us to derive the SFR from the Lyα luminosity. It is however a poor SFR tracer, and measurements are systematically lower than these predictions.

The path length of Lyα photons in the ISM is extremely large due to resonant scattering by hydrogen atoms. Thus, they have increased chances to be absorbed by even a modest amount of dust. As a result, the resultant Lyα emission is quickly decoupled from star formation (Giavalisco et al. 1996). A Lyα escape fraction f (fraction escaping the galaxy and measurable by observers) can be defined by comparing the ratio of Lyα to other SFR tracers (Hα, Hβ, UV). Following such methods, $f \sim 0.1$ is found in the nearby universe and at redshift ~0.3 (Deharveng et al. 2008). The escape fraction seems to increase with redshift, with values of $f \sim 0.2$–0.3 around redshift 2–3, and $f \sim 1$ at very large redshifts (Gronwall et al. 2007; Cassata et al. 2011).

This evolution, statistical in nature, may tell us about the evolution of the dust and of the ISM properties with redshift but still relies on few samples. These results are based on other tracers to measure the actual SFR, and the escape fraction is deduced by comparison. As a conclusion, Lyα is more a tool (to study the statistical properties of the ISM or to detect objects) than a quantitative SFR tracer.

3.4 The Role of Dust

The primary SFR tracers (recombination lines, UV emission) are based on luminosities which are (in most cases) sensitive to the attenuation of light by interstellar dust. The energy "lost" in the UV star light is found again in the mid and far infrared where the emission of the dust grains can be measured. In this section, corrections of primary tracers for the effect of dust attenuation are discussed, as well as the use of the dust emission itself as a SFR tracer.

3.4.1 Extinction Corrections from Recombination Lines

For Hα (or other recombination lines), corrections are usually estimated from line ratios. Observed line ratios are compared to the predictions of case B recombination, and the difference is ascribed to the differential extinction between the two wavelengths. Combined with an extinction curve, it is then possible to estimate the amount of attenuation of, e.g., Hα: A(Hα). Typical attenuations are found between 0 and 2 mag (e.g., Kennicutt 1998a). Another method consists in using the thermal radio to Hα ratio (e.g., Bell 2003).

In most cases, the Balmer ratio is used, for which Hα/Hβ = 2.847 in the T = 10,000 K case B. This view is however challenged by the models of Charlot and Longhetti (2001). When such models are used to simultaneously fit extinction and various line intensities, a dependence on the metallicity of the intrinsic Hα/Hβ ratio is found, which affects the derived extinctions (Brinchmann et al. 2004; Gilbank et al. 2010). This dependence is due to the fact that metallicity affects the cooling and thus the electron temperature of the HII regions. The typical value of 2.847 is given for T = 10,000 K, but the ratio does depend on the temperature (Osterbrock and Ferland 2006).

Observational uncertainties also affect the Hα/Hβ ratio, especially if narrow-band imaging has been used to obtain the data. As mentioned in the Sect. 3.3.2, the Hα has to be corrected for the neighboring [NII] lines. Also, the Balmer absorption features in the underlying stellar spectra have to be corrected for. In the absence of good spectroscopic data allowing to measure it, a standard absorption of 2–5 Å in equivalent width is often used (e.g., Boselli et al. 2001).

3.4.2 Extinction Corrections in the UV

The UV attenuation can be large (reaching over 4 mag). The energy lost by the absorption of UV (and optical) photons is reemitted in the mid and far infrared (between 8 and 1,000 μm). Usually, the total luminosity emited in this range is estimated on the basis of a few available bands (e.g., the IRAS 60 and 100 μm in the past, the Spitzer and Herschel bands in the more modern time). Here, this total infrared luminosity is generically called L(FIR) although different definitions (extrapolating to different infrared wavelength range and using different observations) exist (e.g., Boquien et al. 2010b, and references therein).

Because of this balance of energy, the UV attenuation is related to the $L(\mathrm{FIR})/L(\mathrm{UV})$ ratio. UV-derived star formation rates can be corrected using this ratio (Cortese et al. 2008). A limitation of the method is that at low SFR, part of the FIR emission can come from the heating by old stars, leading to an overestimate of the extinction (e.g., Iglesias-Páramo et al. 2006). Cortese et al. (2008) proposed a method correcting for this effect by using various color indexes to trace the balance between old and young stars.

In many cases, it is difficult to have far infrared data, especially at high redshift. A popular method (although uncertain) in that case consists in using the slope of the spectrum in the UV (β), which is empirically related to the $L(\mathrm{FIR})/L(\mathrm{UV})$ ratio (in the absence of dust, the UV slope is flat, the observed slope, by comparison, can be ascribed to dust). The method is however known to suffer some problems. For instance, different attenuation curves may affect the shape of the spectrum, independently of the amount of dust itself. And mostly, it was shown that local normal galaxies and starbursts provide quite different relations between β and the UV attenuation (Buat et al. 2010; Muñoz-Mateos et al. 2009; Seibert et al. 2005; Kong et al. 2004; Meurer et al. 1999).

3.4.3 Other Considerations on the Dust Attenuation

In the absence of a better indication, "standard" extinctions are sometimes applied to large samples. However, the extinction may depend on the SFR itself (Hopkins et al. 2001; Bell 2003) or on other properties such as the galactic mass or metallicity. While the trends in the nearby universe are relatively well studied, such relationships at high redshift are very uncertain, mostly in reason of the very different selection criteria of various samples (e.g., Gilbank et al. 2010). It is believed that moving toward higher redshift, the extinction increases (as the cosmic SFR does). Eventually, the extinction should decline again when a metal-poor era is reached, and there is some claim of indirect measurements of this effect in the recent high-redshift studies (e.g., Cucciati et al. 2012). The question of the cosmic evolution of dust extinction is however still open and debated.

Another point to note is that extinction corrections are more difficult to apply on small spatial scales than on the global scales of galaxies. For instance, in the case of the UV attenuation estimated from the far-infrared radiation, the heating of a small region can come from UV photons emitted in a neighboring region rather than from the one studied.

The UV emission coming from older stars than the Hα emission, the position of the various type of stars may be decoupled, and their distribution with respect to dust different. In that case, attenuation derived for a population may not apply to another one. In fact, it is thought that the nebular lines are attenuated by roughly twice as much dust as the stellar continuum (because young stars are found inside dust-rich molecular clouds, while the old stellar population has drifted away from the dusty regions (Bell 2003; Calzetti et al. 1994) or because the dust clouds have themselves a finite lifetime (Charlot and Fall 2000). In her "recipe for reddening," Calzetti (1997) suggests that the reddening of the stellar continuum is 0.44× the reddening in the Balmer lines.

3.4.4 Far-Infrared Star Formation Rates

Since a large amount of the stellar emission in the UV is absorbed by dust and the energy is radiated in the far infrared, this dust emission itself can be used as a SFR tracer.

Table 3-2

Secondary far-infrared SFR tracers. The timescale is about the one for the production of UV photons since they dominate the heating of the dust. It however depends on the star formation history because of the contribution to the dust heating by older stars. Calibrations published by various authors using the same IMF as in the primary tracer table are provided. These tracers were calibrated on primary tracers corrected for extinction, mixed tracers, or on models including absorption and emission by dust

Tracer	$\psi(\mathrm{M}_\odot\ \mathrm{year}^{-1}) =$	Comments
FIR	$1.07\,10^{-10}(\mathrm{L(FIR)/L}_\odot)$	Assuming total opacity in the UV (Buat et al. 2008), that works well for high SFRs. Limits: contribution to the heating by old stars, not taking into account the part of the SFR not extinguished
24 µm	$2.50\,10^{-43}(\mathrm{L(24)/erg\,s}-1)$	Calzetti et al. (2010) and references therein.
70 µm	$5.88\,10^{-44}(\mathrm{L(70)/erg\,s}^{-1})$	Calzetti et al. (2010)
160 µm	$1.43\,10^{-43}(\mathrm{L(160)/erg\,s}^{-1})$	This is not a proper calibration because of the large dispersion and the dust heating by old stars according to Calzetti et al. (2010)
250 µm	$8.71\,10^{-45}(\mathrm{L(250)/erg\,s}^{-1})^{1.03}$	From Verley et al. (2010b) but only calibrated in HII regions of M33. Same difficulties as above

Buat and Xu (1996) found that galaxies follow a relation of the type $L(\mathrm{FIR}) = 1{,}680\,\text{Å} \times L(2{,}000\,\text{Å})$ where L(2,000 Å) is the UV luminosity, corrected for extinction, and L(FIR) is an estimate of the total luminosity emitted in the FIR range. Combining such a relation to a calibration between the UV emission and the SFR, one can calibrate a relation between the SFR and the FIR emission (Buat and Xu 1996; Kennicutt 1998a) (➜ *Table 3-2*). This relation is however dispersed by a factor 3 (Buat and Xu 1996), and the method should be valid in galaxies with high star formation rates where dust is mostly heated by young stars, with a negligible contribution by older stars.

Under the assumption that the ISM is totally opaque, the whole luminosity emitted by the stars should be found in the infrared. It is then possible to compute the infrared luminosity L(FIR) for a constant SFR. Buat et al. (2008) proposed for the same IMF adopted above:

$$\log(\psi[\mathrm{M}_\odot\ \mathrm{year}^{-1}]) = \log(L(\mathrm{FIR})[L_\odot]) - 9.97. \tag{3.27}$$

This relation cannot be applied blindly to, e.g., dwarf galaxies and system with low SFRs that are usually suffering low extinctions and in systems with low activity in which the dust is heated by lower mass stars, with longer timescales (but, e.g., Bell 2003, proposed a calibration taking into account this effect.)

Instead of using the total far-infrared emission as above, some authors have argued that it is best to use monochromatic far-infrared indicators. The main reason is that it avoids the need to extrapolate from few measured points to compute the total luminosity and that some wavelengths may be associated in a closer way to the hot dust directly tracing star formation. Calzetti et al. (2009) reviewed many calibrations proposed for the Spitzer telescope observations at 24 µm and proposed relationships for the monochromatic luminosities at 70 and 160 µm. The 8 µm emission was also studied but this wavelength is a poor tracer of star formation (Calzetti et al. 2007). Boquien et al. (2010a) provided a SFR surface density–surface brightness calibration for the Herschel PACS bands at 100 and 160 µm on the basis of Herschel observations of M33.

Herschel SPIRE 250 μm observations of M33 were used by Verley et al. (2010b) to provide a calibration of the SFR from 250 μm.

3.4.5 The Mixed Tracers

Finally, it was recently proposed a family of "mixed" tracers, combining optical or UV, and infrared observables.

As mentioned above, the SFR deduced from the infrared is either empirically derived or is derived under the assumption of full transfer of the energy from the UV to the infrared. For galaxies of low masses, this assumption is usually wrong. It has been proposed to compute the SFR in a more general situation by combining an infrared tracer (that gives us the amount of energy transferred to the infrared by dust), and an optical/UV tracer (that gives us the amount of young stars whose light is not extinguished by dust. Iglesias-Páramo et al. (2006) proposed such a mixed tracer:

$$\psi = \psi_{\mathrm{UV,obs}} + (1 - \eta)\psi_{\mathrm{dust}}. \tag{3.28}$$

The first term is the SFR derived from the observed UV luminosity (following (➲ 3.24)), the second term is the SFR traced by dust (➲ 3.27). η accounts for the IR cirrus emission: this diffuse component of dust heated by the interstellar radiation field generated by old stars, and not directly tracing the SFR. The value of η depends on the samples under study. From 0 for starbursts, it can reach 0.4 for star-forming galaxies (Iglesias-Páramo et al. 2006; Hirashita et al. 2003; Bell 2003).

Kennicutt et al. (2009) proposed several calibrations of mixed tracers composed of Hα (or [OII]) lines and IR (or radio) tracers. They take the form:

$$\psi = C(L(\mathrm{line, observed}) + a\, L(\mathrm{dust})). \tag{3.29}$$

C is the calibration factor that relates directly the luminosity in a line and the SFR. $L(\mathrm{dust})$ is the luminosity in a tracer sensible to the dust-extinguished SFR. Kennicutt et al. (2009) provided values of a when 8 and 24 μm, total infrared, or 1.4 GHz luminosities are used. These values are determined by fitting the coefficients using as a reference the SFR deduced from the Hα line corrected for extinction using recombination lines ratio. Expressed under the form of (➲ 3.29), the $a\, L(\mathrm{dust})$ term is nothing else than a statistical extinction correction. In order to study the spatial distribution of star formation within nearby galaxies, Bigiel et al. (2008) defined a mixed tracer by a combination of 24 μm (from Spitzer) and UV (from GALEX) luminosities since these two wavelength are observed with a similar resolution.

A nonexhaustive list of mixed tracers that have been proposed in the literature can be found in ➲ *Table 3-3*.

3.5 Other Spectral Diagnostics

3.5.1 [OII] 3,727 Å Forbidden Line

Especially at high redshift, the oxygen emission doublet [OII]λ3727 is often used because of its strength to detect galaxies and attempt to measure their SFR. Such forbidden lines are however not as directly related to the stellar emission as the primary tracers. The line luminosity

Table 3-3

Mixed tracers. They are in general calibrated on extinction-corrected primary tracers

$\psi =$	Interest and difficulties
$\psi_{\mathrm{UV,obs}} + (1 - \eta)\psi_{\mathrm{dust}}$	η depends on the sample (Iglesias-Páramo et al. 2006).
$\mathrm{C}(\mathrm{L}(\mathrm{line, observed}) + \mathrm{a\,L}(\mathrm{dust}))$	a L(dust) is a statistical extinction correction, calibrated for 8, 24 μm, total infrared, and 1.4 GHz. Coefficients are given in Kennicutt et al. (2009) for the Hα and [OII] lines.
$0.68\,10^{-28}\mathrm{L}(\mathrm{FUV}) + 2.14\,10^{-42}\mathrm{L}(24\,\mu\mathrm{m})$	Appendix D of Leroy et al. (2008)

is much more sensitive to temperature and ionization state of the gas than recombination lines such as the Balmer lines. The [OII] line has been calibrated empirically by noticing that a good correlation exist anyway between the [OII] and Balmer lines luminosity. Gallagher et al. (1989) found in nearby normal galaxies $L([OII]) = 3.2L(H\beta)$ with however a factor 3 scatter at a given Hβ flux. Kennicutt (1998a) also mentions a large scatter (up to 1 dex) for the [OII]/Hα ratio.

A proper calibration has to take into account the respective extinction in the two lines. Kewley et al. (2004) found an average ratio of [OII]/Hα = 1.2 ± 0.3 that can be combined with a Hα-calibrated SFR to compute the SFR. However, this oxygen line is clearly affected by metallicity (see Sect. 3.8.1).

3.5.2 [CII] 158 μm Fine structure line

The [CII] 158 μm line is the more important coolant of the warm neutral interstellar medium, while the heating of the gas is related to the incident FUV emission (Malhotra et al. 2001, and references therein). It is thus linked to star formation in an indirect way. Malhotra et al. (2001) presented a statistical comparison of the line emission with the far-infrared one: the ratio L([CII])/L(FIR) has a value about $3\,10^{-3}$ with a large dispersion (factor 50). They also found a trend with the temperature of the dust, indicated by the ratio of the luminosities at 60 and 100 μm. Within M31, Rodriguez-Fernandez et al. (2006) showed that the [CII] line traces the regions of star formation, identified with Hα or at 24 μm.

While opening a possible large area of interest for the future, this line is not yet a popular tracer of star formation.

3.6 Radio Emission

The radio emission in normal galaxies come from synchrotron radiation from relativistic electrons and free-free emission from HII regions (see the review of Condon 1992, for all this section). The relativistic electrons are thought to have their sources in the supernovae remnants (SNRs) appearing in galaxies after the explosions of massive stars. The radio emission is thus associated to short-lived massive stars. Since the FIR emission is also linked to star formation (see Sect. 3.4.4), a relation between the radio emission and the FIR emission should be observed, and it is indeed the case in nearby galaxies The FIR-radio correlation can be expressed as (Condon 1992, and references therein)

$$\log(S_\nu) = \log(\mathrm{FIR}) - q. \tag{3.30}$$

q is observed to have a median value of 2.3 at 1.4 GHz (even if the relation is not perfectly linear, especially at low luminosity where the radio emission is lower than predicted by this equation). Though this relation, the radio emission can be considered as a secondary tracer of star formation. Simple models allow to reproduce this relation between SFR, FIR, and radio emission, although the physics is not clearly understood (see Condon 1992, for a much longer discussion). Bell (2003) noted that both the FIR and radio emission underestimate the SFR at low luminosities and proposed a calibration taking this effect into account.

Because the electrons travel in the galaxy before emitting their radiation, this indicator may not bring spatial information on very small scales. Murgia et al. (2002) however suggests that the Hα and radio continuum trace each other well even on scales of a few kpc.

3.7 X-Ray Luminosity

Star-forming galaxies host a number of sources of X-ray emission: high mass X-ray binaries, supernovae remnants and hot plasma (Ranalli et al. 2003). Observations show that the X-ray luminosity (Ranalli et al. 2003) and the number of high mass X-ray binaries (Grimm et al. 2003) correlate with other SFR tracers, what allows to derive the SFR from the X-ray luminosity. The relations provided by Grimm et al. (2003) are included in Table 3-4. The dispersion obtained in the various relations is typically 0.25–0.3 dex. It is also possible to develop more complex models linking this secondary indicator to the SFR (Grimm et al. 2003).

Table 3-4

Multiwavelength secondary SFR tracers. The relations with other tracers are provided instead of actual SFR calibrations (that can be obtained by combining these relationships with the adequate calibration)

Tracer	Calibration	Comments
[OII]	$[\mathrm{OII}]/\mathrm{H}\alpha = 1.2$	Strong at high redshift. Limits: extinction, dependence on ionization state and metallicity.
[CII] 158 μm	$\mathrm{L}([\mathrm{CII}])/\mathrm{L}(\mathrm{FIR}) \sim 3\,10^{-3}$	Large scatter, dependence on dust temperature (Malhotra et al. 2001)
Radio emission	$\log\left(\frac{S_{1.4\,\mathrm{GHz}}}{\mathrm{Wm^{-2}Hz^{-1}}}\right) = \log\left(\frac{\mathrm{L(FIR)}}{3.75\,10^{12}\,\mathrm{Wm^{-2}}}\right) - 2.3$	From Condon (1992). Not affected by extinction. Possible loss of the spatial information
Soft X-ray	$\log\left(\frac{\mathrm{L(0.5{-}2\,keV)}}{\mathrm{erg\,s^{-1}}}\right) = \log\left(\frac{\mathrm{FIR}}{\mathrm{erg\,s^{-1}}}\right) - 3.68$ $\log\left(\frac{\mathrm{L(0.5{-}2\,keV)}}{\mathrm{erg\,s^{-1}}}\right) = \log\left(\frac{S_{1.4\,\mathrm{GHz}}}{\mathrm{erg\,s^{-1}\,Hz^{-1}}}\right) + 11.08$	Ranalli et al. (2003). This tracer has the advantage of not suffering extinction, but the calibrations are dispersed
Hard X-ray	$\log\left(\frac{\mathrm{L(2{-}10\,keV)}}{\mathrm{erg\,s^{-1}}}\right) = \log\left(\frac{\mathrm{FIR}}{\mathrm{erg\,s^{-1}}}\right) - 3.62$ $\log\left(\frac{\mathrm{L(2{-}10\,keV)}}{\mathrm{erg\,s^{-1}}}\right) = \log\left(\frac{S_{1.4\,\mathrm{GHz}}}{\mathrm{erg\,s^{-1}\,Hz^{-1}}}\right) + 11.13$	

3.8 Additional Factors

All the tracers have been discussed above adopting an IMF and assuming a solar metallicity. Adopting other assumptions could change the calibrations of the SFR tracers easily by a factor up to 2 (see below), even for the primary tracers. Another difficulty is the steady-state assumption that may not always hold. These aspects are discussed in the rest of the section.

3.8.1 The Effect of the Metallicity

The primary tracers of the SFR depend on the metallicity as low-metallicity stars have higher temperatures, higher UV luminosities and ionizing fluxes. Bicker and Fritze-v. Alvensleben (2005) calibrated the dependence on the metallicity of the C factors in (➲ 3.24) and (➲ 3.26). Both calibration factors vary with the metallicity by a factor about 2 between low and high metallicity. The reader is refereed to Bicker and Fritze-v. Alvensleben (2005) for quantitative values.

The [OII] line emission can also be computed through detailed modeling, and Bicker and Fritze-v. Alvensleben (2005) provide also a calibration of its dependence on metallicity. The influence of metallicity on the luminosity of this line was also investigated by empirical studies that established that the [OII]/Hα ratio does vary with metallicity. This dependence was calibrated by Kewley et al. (2004). Mouhcine et al. (2005) however showed that even this calibration does not catch all the physics at play. They found a non-monotonic relation between the [OII]/Hα ratio and the metallicity. They also found that, in addition, the ionization state affects the ratio, especially in metal-poor active starbursts.

Technically, all the secondary tracers calibrated on one of the primary tracers (with an explicit dependence on metallicity) should present the same dependence on metallicity.

3.8.2 Choice of the IMF

Since the primary tracers are related to the more massive stars, but the SFR is integrated over the whole IMF, the calibration factors are very sensitive to the form of the IMF (that determines how many massive stars are present per unit stellar mass formed). A proper computation is necessary to provide a calibration. ➲ *Table 3-5* quotes a few transformation formulas that can be found in the literature and provides an idea of the uncertainty due to the IMF. Seeing these

■ **Table 3-5**

Effects of the choice of the IMF on the SFR calibrations. This nonexhaustive list of values is collected from the literature (they are not 100% consistent with each other)

Conversion	Reference
ψ(Kroupa 2001) × 1.5 = ψ(Salpeter)	Brinchmann et al. (2004)
ψ(Kroupa 2001) × 1.59 = ψ(truncated Salpeter)	Bigiel et al. (2008)
ψ(Chabrier 2003) × 1.5 = ψ(Salpeter)	Schiminovich et al. (2010)
ψ(Kroupa 2001) × 1.5 = ψ(Salpeter)	Argence and Lamareille (2009)
ψ(Kroupa 2001) × 0.88 = ψ(Chabrier 2003)	Argence and Lamareille (2009)

numbers, it is obvious that the choice of the IMF can affect the SFR by a factor at least as large as the metallicity influence or the fraction of Lyman continuum photons not contributing to the ionization of the gas.

There is good reason to think that the IMF is universal (Bastian et al. 2010). In that case, such a global correction should allow us to switch from one IMF to another. However, the idea of variations of the IMF come back to haunt extragalactic astronomers frequently. GALEX allowed discovery of a surprisingly high fraction of galaxies with extended UV emission, interpreted as extended star formation (see ➋ Sect. 4.3). This discovery, together with the fact that Hα (very sensible on the IMF slope at high masses) had not revealed this star formation, previously led to some speculations about the possible variation of the IMF with radius within galaxies, probably through a physical dependence on other quantity, such as the gas density or the total SFR. Krumholz and McKee (2008) proposed a gas density threshold for the formation of massive stars, while Pflamm-Altenburg et al. (2007) suggested that the IMF for a whole galaxy (what they call the integrated galactic initial mass function, IGIMF) results from the combination of the cluster mass distribution and the IMF within clusters. In their approach, the maximum stellar mass in a star cluster is limited by the cluster mass, and the maximum cluster mass depends on the SFR. As a result, they found that the IGIMF does depend on the SFR. If confirmed, such a variation would have important consequences, the first one being that the calibrations discussed above would be false, especially at low SFRs.

This motivated several studies on measurements of the Hα/UV ratio in small galaxies and outer parts of spirals and its interpretation, with conflicting results. In addition to the difficulty due to measurements and calibrations, the situation is complicated by the fact that this ratio may also depend on other factors (e.g., different attenuations, variation of the fraction of ionizing photons absorbed by the dust, micro-history of the star formation: see next section). As a result, a clear situation has not emerged yet, but several works are ongoing on the subject.

A variation of the IMF with the redshift has also been advocated (with a top-heavy IMF at high redshift), for instance, to reconcile the evolution of the SFR and stellar mass cosmic average densities (e.g., Wilkins et al. 2008a).

In summary, suggestions of IMF variations are proposed in various contexts, and their possibility should be kept in mind. However, the subject stays under debate because of the complexity of the observations and the possible alternate explanations.

3.8.3 Effect of the Star Formation Micro-History

The steady state assumption was applied to calibrate the relationships between star formation tracers and SFR described before. This seems to be a reasonable assumption when considering normal galaxies as a whole. However, it cannot be applied to, e.g., very small individual regions inside galaxies (e.g., individual star clusters). A minimum scale is needed to average out the stochasticity of star formation. The assumption may also break down for starbursts (or post-starbursts), in which the star formation history is undergoing (or has undergone) large and sudden variations, or in small galaxies, where for stochastic reasons the micro-history of star formation (over the last 10–100 Myr) may present significant variations. When there is reason to believe that the star formation history was not constant over δt, time corresponding to the timescale of the tracer used, a meaningful SFR cannot be derived with this tracer. It is still possible to study these objects, but it is then necessary to obtain data constraining, e.g., the total mass that was formed during a star-forming event together with the age and duration of this event.

Using resolved stars, McQuinn et al. (2010, and references therein) found that the duration of starbursts in nearby dwarf galaxies is several 100 Myr (longer than many other estimations of a starburst duration), allowing to derive reasonable SFR for tracers with similar timescales (e.g., UV). But they also found that Hα measurements sometimes provide a different result because the SFR fluctuates on a few Myr timescale. Boissier et al. (2008) also suggested that the colors of their LSB galaxies may be due to a succession of active star formation and quiescent phases.

The effect of such bursts and micro-star formation history on SFR tracers (and their ratio, especially Hα/UV) has been much discussed in, e.g., Weilbacher and Fritze-v. Alvensleben (2001), Iglesias-Páramo et al. (2004), Boquien et al. (2007), and Boselli et al. (2009). It is however difficult to disentangle from other possible effects, e.g., the IMF variations discussed in the previous section.

4 Star Formation Observed in Galaxies

Provided the assumption and limitations discussed in ➔ Sect. 3, it is possible to compare the SFR actually observed to the theoretical ideas presented in ➔ Sect. 2 and more generally discuss some empirical properties of star formation on galactic scales.

4.1 Star Formation in the Local Galaxies

Surveys of the local volume show that star formation in galaxies depends on their luminosity, gaseous content, or morphological type. This can be seen in ➔ *Fig. 3-3* reproduced from Karachentsev and Kaisin (2010) who combined Hα observations from galaxies within 8 Mpc (in addition, the overplotted open square indicates typical values for the Milky Way). The SFR is also found to depend on other quantities, such as colors (as more luminous galaxies are also redder). Another Hα survey of the local 11 Mpc can be found in Kennicutt et al. (2008). A similar study based on the ultraviolet tracer of star formation (instead of Hα) can be found in Lee et al. (2009). Some differences between Hα and UV are found especially in the smallest galaxies; see ➔ Sect. 3.8.2 for a possible interpretation. Many other surveys in, e.g., *Hα*, UV, and FIR provide different views according to their selection criteria (as discussed in the review of Kennicutt 1998a).

The SFR range from $\sim$0 in early type galaxies to a few M$_\odot$ year^{-1} in the Milky Way and relatively normal disk galaxies. However, much higher values of the SFR (up to several 100 M$_\odot$ year^{-1}) are also found in rarer starbursts (absent in the small local volume), usually associated with signs of interaction or mergers.

4.2 The Schmidt Laws

4.2.1 Preliminary Considerations: Many "Laws"

An important part of the study of star formation on the scales of galaxies revolves around the so called Schmidt Law. Although now largely in usage, the term "law" is not really appropriate as

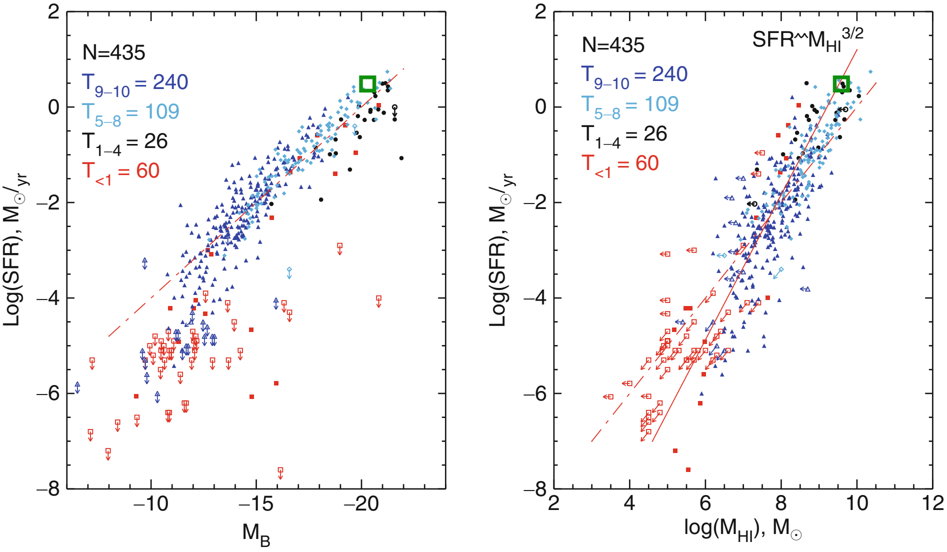


Fig. 3-3

Reproduced by permission of the AAS from Karachentsev and Kaisin (2010): Hα derived SFR in the Local Volume as a function of the B magnitude (*left*) and HI mass (*right*). The *dashed lines* correspond to a constant ratio between y- and x-axis . The *open square* indicates values typical for our Milky Way (e.g., Boissier and Prantzos 1999; van den Bergh 1999, and references therein)

it is really an empirical relationship, originally established between the number of young stars in the Milky Way and the gas volume density by Schmidt (1959). This law can be writen in this form:

$$\rho_\psi \propto \rho_{\text{gas}}^n. \qquad (3.31)$$

A lot of relations similar to (3.31) are called "Schmidt Laws." The first variation is the use of surface densities rather than volume densities, suggested by Sanduleak (1969) in his study of the Schmidt Law in the Small Magellanic Cloud, a form that is very frequently used:

$$\Sigma_\psi \propto \Sigma_{\text{gas}}^n. \qquad (3.32)$$

This form has the advantage to be more easy to reach from observations (no need to know the scale heights of disks). For a constant scale height, the index n of the volume law (3.31) is obviously the same as the one of the surface densities law (3.32). Surface densities also appear in many of the equations of Sect. 2, suggested by various theories. Indeed, surface densities in a rotating disk do play a physical role. Such a local relation between surface densities of star formation rate and gas is called "Schmidt-Sanduleak Law" by Madore (2010).

Madore et al. (1974) noticed that the index of the Schmidt Law was different in the inner and outer parts of M33, suggesting that the star formation law may change with radius within galaxies. Later, the "radial" Schmidt Law was studied on the basis of azimuthally averaged profiles (Kennicutt 1989; Kennicutt 1998b; Wong and Blitz 2002; Boissier et al. 2003, 2007).

Finally, the Schmidt Law was also studied on the scale of whole galaxies, for instance, by computing average surface densities of gas and SFR, as it is done in the seminal paper of Kennicutt (1998b). Sometimes, the total SFR and total gas amount are directly compared one to another (without normalization per, e.g., size). Following again Madore (2010), such a global law (normalized or not) can be called a "Schmidt-Kennicutt Law."

Very often, these various types of laws are compared one to another or plotted together even if this should be done only with a lot of caution. For instance, if star formation follows a local Schmidt Law, with a local critical density threshold, an analysis of the radial Schmidt Law may provide a steeper slope, and the threshold may be hard to recover (see ➲ *Fig. 3-4*). Similar differences will exist between the local Schmidt Law and a measurement of the Schmidt-Kennicutt Global Law (➲ *Fig. 3-5*). The illustrations of these differences (➲ *Figs. 3-4* and ➲ *3-5*) are qualitative, but based on actual simple toy models. Leroy et al. (2008) proceeded to a similar exercise: they implemented a pixel-by-pixel threshold in their data. When computing radially averaged profiles, they found that the threshold damps the average SFR but does not bring it to 0. Kennicutt (1989) had already noticed that the globally averaged star formation rate is more strongly coupled to the HI density than the radial profile, illustrating again that the various

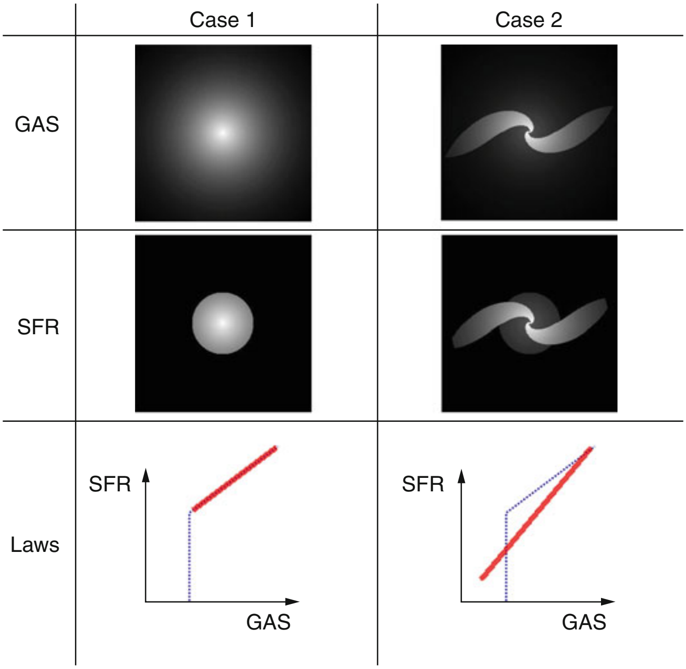


Fig. 3-4
This sketch illustrates the possible difference between a local law (*dotted blue*) and an azimuthal law (*solid red*). In case 1, the disk is purely exponential and there is no difference between the local and azimuthal laws. In case 2, two spiral arms enhance locally the density by a factor 3. While this toy model implement the exact same local law as the previous one, its slope and threshold are not recovered in an azimuthal analysis. The differences between the local and azimuthal laws are sensitive to the covering fraction of the arms and its variation with radius

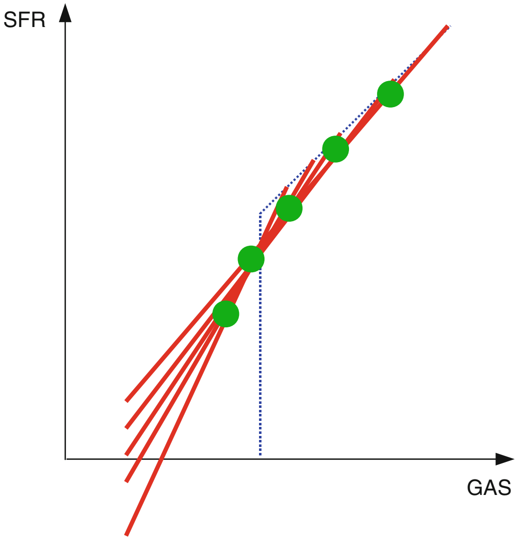


Fig. 3-5
In this illustration, several toy models were constructed with the same local law (*thin blue dotted line*). They each follow different azimuthal laws (*thick solid lines*) depending on, e.g., the radial variation of the spiral arms covering fractions. The *green points* show the global Schmidt-Kennicutt Law. Local, azimuthal, and global relationships have different slopes and do not show the same threshold

scales follow intrinsically different laws. These variations may partly explain the large range of values found in the literature for the index n (0–4), although part of this scatter traces actual differences (Kennicutt 1998a) and the limitations of a simple Schmidt Law parametrization (Kennicutt 1997), in addition to the various SFR tracers used.

These variations (local, azimuthal, global) are not the only one. The results may also depend on the SFR tracer used (see Sect. 3), but also on the gas phase used. "Gas" refer in this chapter to the total gas, but often only the neutral (mostly HI) or molecular (mostly H_2) is used[2] (sometimes for observational reasons, sometimes because one wants to study a process specific to one phase). In most cases, H2 is actually not observable, and other molecules are used to trace it. The most famous one is CO, but the X_{CO} conversion factor carries large uncertainties (see discussion in Calzetti and Kennicutt 2009). While some authors use a metallicity dependence, other prefer to keep a constant value, what may directly affect the slope of the Schmidt Law.

Table 3-6 attempts to summarize the various type of laws and the phenomenon that they probe.

Finally, the Schmidt Laws have been extended to include other factors than the gas density, especially to test the various theoretical expectations discussed in Sect. 2 (for instance, 3.23). This was motivated by the theoretical expectations, but also by the fact that a perfect fit cannot

[2] To account for the abundance of the other elements (mostly Helium), a corrective factor ~1.36 is sometimes used to go from HI to neutral and from H_2 to the molecular contents. On the other hand, some authors call "total gas" the sum of the HI and H_2 masses, i.e., the total mass of hydrogen, not including this factor.

Table 3-6

Proposed relevance of the various Schmidt Laws. Secondary factors not included!

	Local Schmidt-Sanduleak	Azimuthal Radial Schmidt Law	Global Schmidt-Kennicutt
HI	Local effects on HI/H2 phases transition	Processes affecting the formation of molecular gas on orbital timescales (e.g., spiral arms)	Transformation of the global reservoir of HI into H2
Total	Local gravitational effects	Gravitational processes occurring on orbital timescales (e.g., role of Ω)	Role of the global reservoir of gas
H2	Formation of stars in GMCs		

be obtained with a pure dependence on the gas density. First, the observed simple Schmidt Law presents a large scatter, believed to be larger than the observational uncertainties (e.g., Kennicutt 1998b). Second, it was noticed in many occasions that the Milky Way and a number of other nearby galaxies present flat HI profiles, dominating the total gas at large radii, while their SFR clearly decreases with radius. A pure dependence on the total or HI surface density alone (at least in azimuthal profiles) cannot work (e.g., Blitz and Rosolowsky 2006; Boissier and Prantzos 1999; Ferguson et al. 1998; Kennicutt 1989).

4.2.2 Which Law Is Right?

Which of the Schmidt Laws described in Sect. 4.2.1 should be used to, e.g., constrain our models ? It actually depends on the precise goal that is aimed at. To understand the very physics of star formation (how gas is turned into stars), local studies are of prime importance (the issue on which scale "local" studies should be performed is discussed below). If one wants to constrain a 2D model of galaxies, then again the local law brings more constraints.

However, it is well known that some properties depend mostly on the galactocentric radius *R* (e.g., the fact that star-forming galaxies display an exponential disk, the abundance, and color gradient). This is related to the fact that these quantities results from processes occurring on timescales longer than the rotation time around the center of the galaxy. 1D models (depending only on radius) are thus performed, especially in studies of chemical evolution. Such models aim at reproducing the radially averaged SF law (that tells us about how star formation is related to the gas on timescales similar to the rotation time).

The total amount of gas is difficult to measure at high redshift, and a "reversed" Schmidt Law is sometimes used to determine the amount of gas from the measured SFR (e.g., Péroux et al. 2011). If the starting observation is a surface brightness, then a local law should be used. On the contrary, a global SFR measurement should be combined with a global Schmidtt-Kennicutt. Studying the global law is however not limited to the remote universe. One can see it also as a study of the rate at which the global reservoir of gas in a galaxy becomes available for star formation. In this case, the Schmidt Law is not really a way to study the process of star formation itself, but also the role of gas accretion or other external processes (e.g., Schiminovich et al. 2010).

4.2.3 Which Scale Is Right?

The question of the smallest scale on which the Schmidt Law should be studied is often asked. Koda (2008) suggests that the properties of GMCs provide two limits. The first one is the "drift scale." Young stars have a relative velocity with respect to the gas clouds. In combination with the typical timescale of the star formation tracer used (10 Myr for Hα, 100 Myr for the UV), one can expect a separation between the SFR tracer and the physically related gas of 100 pc–1 kpc. At smaller scales, the SFR tracer and the underlying gas are not physically related. The second scale advocated by Koda (2008) is the separation between GMCs, which is typically 200 pc in the Milky Way, and provide an order of magnitude for other galaxies. The "statistical" behavior of star formation can be seen only if the adopted typical resolution includes at least a few GMCs. Some of the theoretical argument discussed in ➤ Sect. 2 also apply on large scales (e.g., gravitational collapse), up to the kpc scale. Thus, also from this point of view, the Schmidt Law should be valid only on scales larger than at least a few 100 pc.

In their simulations, Feldmann et al. (2011) found that the scatter in the SFR–molecular gas relationship increases rapidly with decreasing averaging scales. In their view, this is due to the fact that the molecular gas is the local "instantaneous" one, while the SFR is time-averaged over 100 Myr timescales.

All these considerations suggest that the Schmidt Law should break down when looking at scales smaller than a few 100 pc and that the scatter should increase when going from the kpc scale downward. Observational work in nearby spirals have recently reached this level (local laws on few 100 pc resolution), and an increase of the scatter of the empirical law when going from the kpc to the 100 pc scales is actually observed (Thilker et al. 2007b; Kennicutt et al. 2007; Bigiel et al. 2008; Verley et al. 2010a; Liu et al. 2011).

4.2.4 Schmidt Laws: Current Observational Status

The left panel of ➤ *Fig. 3-6*, taken from Kennicutt (1998b), shows the Schmidt-Kennicutt Law, with a best fit relation often used in the literature:

$$\Sigma_\psi = 2.5\,10^{-4}\left(\frac{\Sigma_{\mathrm{gas}}}{1\mathrm{M}_\odot\,\mathrm{pc}^{-2}}\right)^{1.4} . \tag{3.33}$$

It is impressive that this relationship holds on 5 orders of magnitudes of the gas density, showing that there is certainly some level of universality in star formation, related to fundamental processes. Also, the relation is very close (index 1.4 vs. 1.5) to the one proposed in the theoretical section on the simple basis of free fall as the main driver for star formation (➤ 3.10). However, it should be noticed that if one focus on one type of galaxies, e.g., normal spirals, then the scatter remains important, and the slope is not so clearly established. Nevertheless, many other studies have found this relation to give a satisfactory description in many different cases: azimuthal profiles of outer disks (Boissier et al. 2007), tidal dwarf galaxies (Lisenfeld et al. 2001), low surface brightness galaxies and dwarfs (may be with some shift downward with respect to the canonical relation; see, e.g., Boissier et al. 2008; Wyder et al. 2009).

The right part of ➤ *Fig. 3-6* shows that a Schmidt Law including a dynamical factor ($\Omega \propto 1/\tau$) suggested on several theoretical grounds (see ➤ Sect. 2) works as well as the simple Schmidt

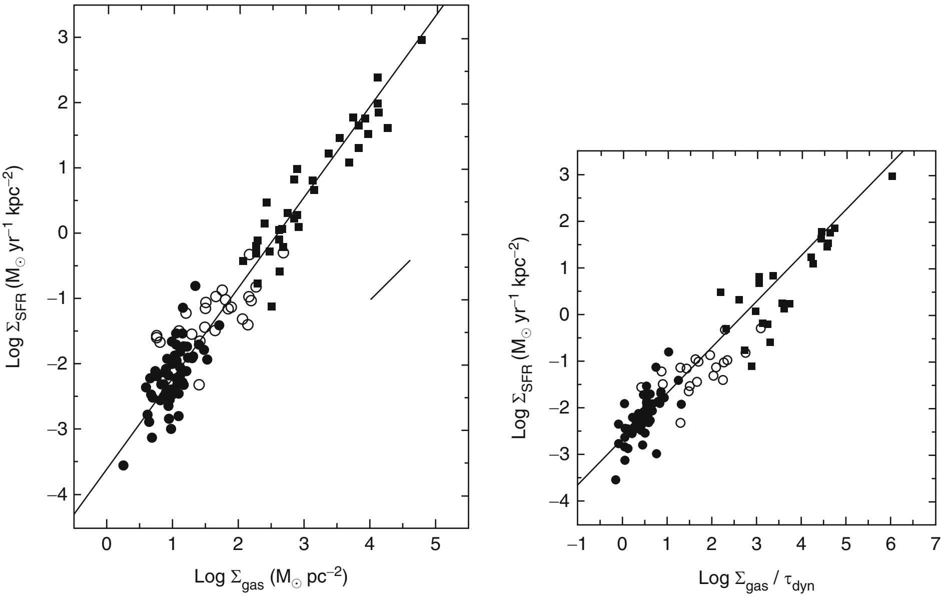


Fig. 3-6

The two panels are reproduced by permission of the AAS from Kennicutt (1998b). On the *left*, the relation between the average SFR and gas surface density is shown. The global Schmidt-Kennicutt Law is valid over 5 orders of magnitudes. On the *right*, a Schmidt Law modulated by a dynamical factor is shown. Both type of laws are in reasonable agreement with the data. The *solid circles* correspond to normal *spirals*, *squares*: circum-nuclear starbursts, *open circles*: central regions of normal disks

Law depending only on the gas density. On this empirical basis, there is no reason to choose one with respect to the other one.[3] The fit by Kennicutt (1998b) for this law is simply

$$\Sigma_\psi = 0.017\Sigma_{\rm gas}\Omega. \tag{3.34}$$

With the advent of high-quality UV data (GALEX satellite) and infrared data (Spitzer), it became recently possible to obtain for large samples of nearby galaxies detailed maps of the SFR (as a combination of UV and far infrared). A noticeable benchmark was obtained by the THINGS team who combined such data with HI and CO maps allowing to proceed to a detailed study of the local Schmidt-Sanduleak Law for a representative sample of spirals and dwarfs (Bigiel et al. 2008; Leroy et al. 2008). Their results are discussed below in quite some details because they represent the state of the art and as such establish a new reference (even if some of their results had been found before on the basis of other samples or methods).

Bigiel et al. (2008) showed that the best local correlation is obtained between the SFR and the molecular gas (see *Fig. 3-7*). Their best fit is

$$\Sigma_\psi = 10^{-2.1\pm0.2}\Sigma_{H_2}^{1\pm0.2}. \tag{3.35}$$

[3] See however the final remark of Sect. 4.2.1

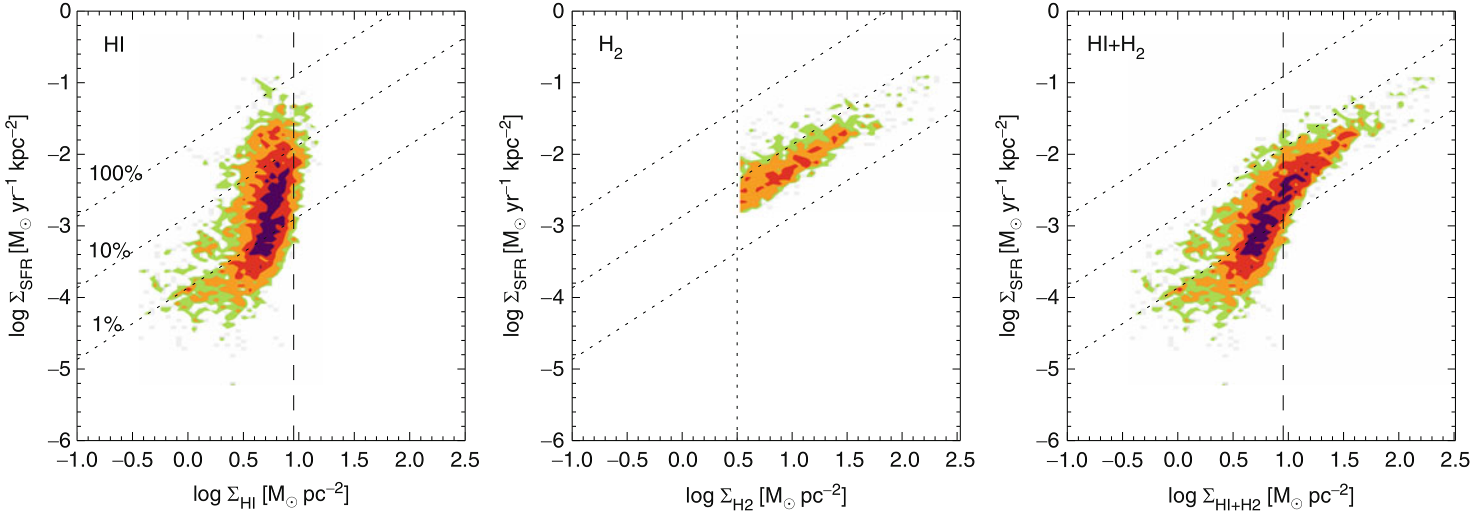


Fig. 3-7

Local Schmidt-Sanduleak Laws observed in Bigiel et al. (2008) for the HI, H2, and total gas (Reproduced with permission of the AAS)

Bigiel et al. (2008) interpret this index $n = 1$ as the fact that several clouds are included within the surface studied at their resolution, and the value of Σ_{H2} is only dictated by the number of clouds within the beam. Then, these clouds all form stars with a universal behavior. There is no clear relation with HI (but a saturation, discussed in ❯ Sect. 4.3). The SFR-total gas relation presents a two-slope behavior. At high density, the relation is dominated by the linear relation between the dominant molecular component and the SFR. At low density, it is dominated by the HI/H2 transition. This relation is reproduced in the simple approach of Madore (2010) presented in ❯ Sect. 2.3.5. It is also well reproduced by the Krumholz et al. (2009) model (❯ 2.3.7), that has been shown to be also in agreement with other observations (e.g., in QSO absorbing systems). This model, however, is purely local (thus it should not be applied to global or azimuthal observations blindly). Krumholz et al. (2009) in fact proposed that, e.g., gravitational instabilities, spiral arms, and other global processes regulate in fact the total gas distribution. Then, their theory predicts the right amount of molecular gas and the SFR.

Leroy et al. (2008) went one step further with respect to the results presented in ❯ *Fig. 3-7*. They tried to fit their data with a selection of several possible laws for the transformation of HI into H2 and stars. Despite extensive tests, they could not definitely pinpoint one unique driver for star formation. Many effects are observed (dependence of the ratio $\Sigma_{\psi}/\Sigma_{\rm gas}$ on the, e.g., radius, orbital timescale, stellar density) but not a single theory can explain the observations over the full range of gas density and in all type of galaxies beyond any doubt. They established that the molecular fraction decreases with radius (what was suggested by previous observations nevertheless) but also with decreasing stellar density, hydrostatic pressure, and orbital timescale (or equivalently Ω). Those dependence are quantified in four scaling relations proposed in their Sect. 5.4.5:

$$\begin{aligned}
\Sigma_{H2}/\Sigma_{HI} &= 10.6\exp(-R/0.21R_{25}) && (3.36)\\
&= \Sigma_{*}/81\mathrm{M}_{\odot}\,\mathrm{pc}^{-2} && (3.37)\\
&= (P/1.7\,10^{4}cm^{-3}Kk_{B})^{0.8} && (3.38)\\
&= (\tau_{\rm orb}/1.8\,10^{8}yr)^{-2}, && (3.39)
\end{aligned}$$

where R_{25} is the isophotal radius at 25 mag arcsec^{2} in the B band and $\tau_{\rm orb}$ is the orbital timescale, thus $\tau_{\rm orb} = 2\pi\Omega^{-1}$.

Note that the dependence on P can be hiding a dependence on the gas density, as already mentioned (see Blitz and Rosolowsky 2006) and that Leroy et al. (2008) did not test relations with "free" index such as (❯ 3.23) but only precise relations predicted by various theories (if several phenomenon contribute to the final relation, one can imagine that the final observed relation could be in between the various predictions).

4.2.5 Starbursts and the Schmidt Law

Most of the discussion above concerns spirals and dwarf galaxies. Kennicutt (1998a) included however in his study the circumnuclear starbursts (see ❯ *Fig. 3-6*) and found an index 1.4 for the Schmidt-Kennicutt Law. How can this value be reconciled with the linear relation of Bigiel et al. (2008) with H2 (at these high densities, most of the gas is molecular)? Gao and Solomon (2004) proposed that the SFR presents in fact a linear relation with the dense gas density (as traced by HCN) rather than with the molecular content (as traced by CO). In "normal galaxies," the ratio of dense to molecular gas (the dense gas fraction) is constant. Thus, a linear relation is obtained between the SFR and CO. In starbursts, the SFR is still proportional to the dense gas, but the dense gas fraction becomes larger so that the index n of the Schmidt Law (with either the molecular or the total gas) will be larger than 1. The enhancement in the dense gas fraction in starbursts with respect to normal galaxies has still to find a physical cause, probably linked to the much higher densities and short timescales that they experience (as mentioned in ➲ Sect. 2.4). Adding the dependence on the dynamical timescale allow starbursts and normal galaxies to follow a similar relationship, even at high redshift (Genzel et al. 2010; Daddi et al. 2010).

4.3 Observed Thresholds

As seen in ➲ Sect. 2.2, various theories for star formation predict the existence of a threshold, or a critical surface density of gas $\Sigma_{\rm crit}$ below which star formation should not proceed. This motivated several empirical studies, such as the early one of Kennicutt (1989) who observed thresholds in the range 1–10 $M_\odot$ pc^{-2}. Other works suggesting the existence of a sharp threshold followed (Kennicutt 1998b; Martin and Kennicutt 2001), mostly based on Hα azimuthal profiles. A number of more recent works based on another star formation tracer (namely, the UV emission) challenged this view, as star formation (UV emission) was found at very large radius in many nearby galaxies. Boissier et al. (2007) showed that most of the UV azimuthal profiles were extending quite smoothly beyond the optical radius, where it was thought previously that star formation was indeed present but only rare and stochastic (e.g., Ferguson et al. 1998). This UV emission at low density was a very generic finding since the GALEX satellite revealed the existence of so-called XUV (eXtended UV) galaxies (Gil de Paz et al. 2005; Thilker et al. 2005). It was realized little after their discovery that this phenomenon concerns a large fraction of disk galaxies (about 30% according to Thilker et al. 2007a).

Measuring the Hα emission at large radii is quite challenging. Goddard et al. (2010) proposed a new analysis of existing Hα and FUV images for 21 galaxies, computing Hα profiles with an enhanced method (addition of detected object fluxed in radial bins) with respect to simple azimuthal averages. This allowed them to make a detailed comparison of FUV and Hα profiles. They classified their objects into "normal" disks (for which a break is observed in both UV and Hα close to the optical radius) representing 50% of their sample. Note however that a break (change of slope) is observed rather than a very sharp truncation (threshold). The other half of their sample consist in UV extended galaxies. Of these, they claimed that 6 out of 10 galaxies are also extended in Hα, and only 4 out of 10 galaxies have a UV smooth profile and a sharp truncation in Hα.

On a local basis (by opposition to the azimuthal profiles discussed above), the recent work by Bigiel et al. (2008) suggests that there is a (local) critical density for HI, above which the gas is mostly molecular. In other words, there is a saturation of Σ_{HI} around 10 $M_\odot$ pc^{-2} (what had been previously suggested, e.g., Blitz and Rosolowsky 2006). However, the analysis of Leroy et al. (2008) cannot ascribe clearly this "critical density" to any of the threshold theories they tested (Toomre and other). In fact, the saturation corresponds to a phase transition from HI to H2: it does not mean that there should be an absolute threshold for star formation in the total gas density. Although the saturation density seems to favor a Schaye (2004) critical density, Leroy et al. (2008) argue that according to this assumption, the full disk should be supercritical. While it allows to predict an edge for the disk, it does not predict locally within a galaxy the regions forming stars or not.

In conclusion, the idea of a threshold radius beyond which star formation is totally suppressed (and even of a critical density) seems to be of little predictive power to decide where star formation proceeds or not. If there is indeed a gravitational instability threshold corresponding to the full disk being instable and prone to form stars, then beyond the corresponding radius, star formation still proceed, may be due to local enhancements of the density, an idea followed in the simulations of Bush et al. (2008).

Elmegreen and Hunter (2006) proposed that this absence of clear threshold shows that star formation is the results of different processes all combining to produce the observed radial trends. In their view, the SFR saturates at a maximum possible rate independently of the responsible physical process. Especially in outer, low-density regions, this allows to sustain a low level of star formation even when some of these possible processes shut down.

4.4 Relations to the Stellar Content: The Specific Star Formation Rate

The possible influence of the stellar surface density on the SFR surface density was discussed in Sect. 2. Such influence has been studied empirically (e.g., Shi et al. 2011; Abramova and Zasov 2008; Leroy et al. 2008; Dopita and Ryder 1994) but is not clearly favored with respect to other theoretical suggestions. The fact is that a relation seems to exist between the SFR and the stellar mass in the nearby universe (Brinchmann et al. 2004) and at high redshift (Boissier et al. 2010, and references therein). The current interpretation is however that this relation is not causal, but reflects the star formation history of the galaxies.

In fact, the SFR to stellar mass ratio, the specific star formation rate (SSFR), has become an important diagnostic tool to study galaxies (e.g., Buat et al. 2008; Noeske et al. 2007; Bell et al. 2005). The SSFR has the advantage over the SFR to be a normalized quantity, representing the balance between the present activity of a galaxy (the SFR) and its past one (the stellar mass accumulated up to this point). A related quantity is the birthrate parameter b, the current SFR to past average SFR ratio (Kennicutt 1998a). Empirically, it has been directly estimated from the equivalent width of the $H\alpha$ line, and it is also related to broadband colors of galaxies. Salim et al. (2005), for instance, provide a relationship between b and the NUV-r color. Nowadays that stellar mass are relatively easy to estimate, the use of SSFR is more frequent than the one of b. They are however directly related (in a closed box model of evolution with a returned fraction R, age T, SSFR = $b/((1-R)T)$.

The SSFR (or b) is thus an extremely interesting quantity to study galaxy evolution. In the nearby universe, there is a relation between the Hα equivalent width (related to b, see Kennicutt 1998a) and the morphological type (shown in *Fig. 3-8* with the data of James et al. 2004). This relation is due to the differences in the average star formation histories of galaxies of different type as illustrated by the sketch of Sandage (1986a). SSFRs were used also in the context of radial profiles to get clues on the growth history of galactic disks (Muñoz-Mateos et al. 2007) or at high redshift to unravel the mysteries of galaxy formation at the earliest epochs (e.g., Noeske et al. 2007).

4.5 Star Formation History

The SSFR was presented in Sect. 4.4. This quantity gives an important clue on the star formation history of galaxies, but only in a rough sense: the present SFR versus the past averaged

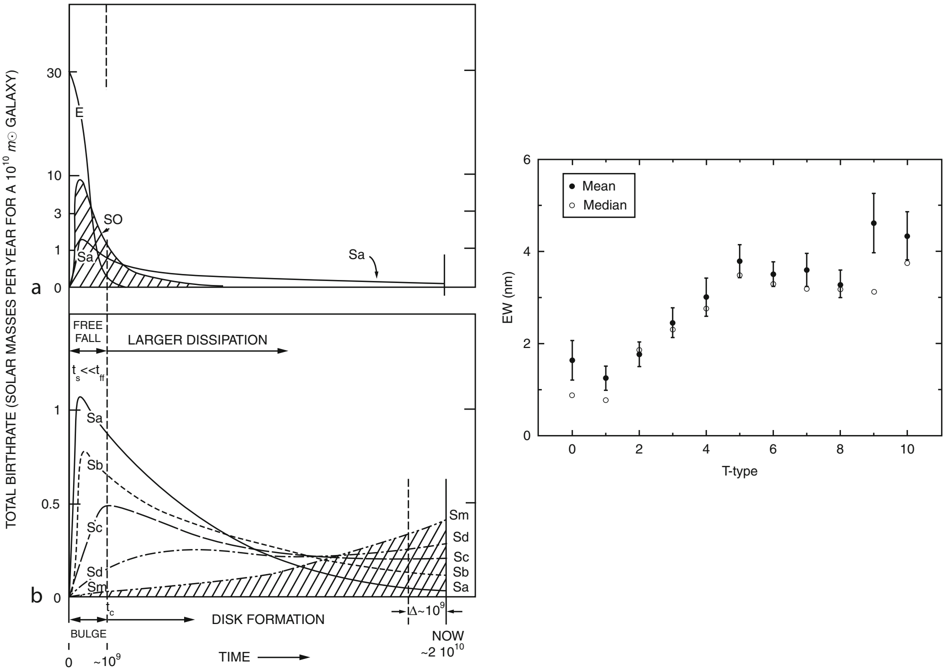


Fig. 3-8
***Left*: Sketch of the average star formation histories of galaxies with various types (taken from Sandage 1986a). Early type are show in panel a and late type galaxies in panel b. Obviously, the *b* parameter (or the SSFR) increases from early to late type. *Right*: Equivalent width of the Hα line (that can be considered as a measure of the birthrate parameter *b*, see Kennicutt 1998a) as a function of type, taken from James et al. (2004) (Credit: Sandage 1986b; James et al. 2004, reproduced with permission ©ESO)**

one. Ideally, to decipher the evolution of galaxies, it is needed to know its star formation history (SFH), i.e. the variation of the SFR with time. In some cases, it is possible to derive it from observations.

An indirect method consist in comparing the predictions of models assuming SFH (e.g., exponentially declining star formation rates or more complex histories) with observational sets. This can be the case for instance of chemical evolution models that are constrained by, e.g., age-metallicity, gas fractions, and metallicity distributions, Spectral synthesis models can be constrained by observed spectra, spectral energy distributions, or simply colors. These indirect methods are not discussed in details here. In the remaining, more direct methods are presented.

In the Milky Way, it is possible to determine the ages of large number of stars. By measuring how many stars are found within bins of ages, it is possible to reconstruct the SFH. One still has to be careful about which component is considered (see, e.g., Wyse 2009): in the thin disk, it is in general assumed (on the basis of observations and chemical evolution models) that the SFR in the solar neighborhood has been oscillating around a more or less constant value, but it should have declined between its early existence and the current epoch if the whole disk is considered. The thick disk and the halo should have different histories according to their formation scenario. Rocha-Pinto et al. (2000) used 552 late-type dwarfs chromospheric ages to derive a SFH consisting in a series of small amplitude bursts in their 0.4 Gyr bins. Fuchs et al. (2009) used a sample of M-Dwarfs to derive an increase of the Milky Way SFR when going backward in time and compared it to several previous determinations.

Another method consists in the analysis of color-magnitude diagrams (CMDs). Combining the observed position of stars in CMDs with stellar evolution models, it is possible to recover the star formation history of galaxies. This approach has been used in the Milky Way (Hernandez et al. 2000; Cignoni et al. 2006). Especially with the observations of the Hubble Space Telescope, it became possible to resolve stars in many galaxies and apply this method to determine their SFHs. It was found that the local group dwarfs present a large variety in their SFH, with extended episodes of activity separated by short quiescent phases (see, e.g., Tolstoy et al. 2009, and references therein). Harris and Zaritsky (2009) used CMDs in the large magellanic cloud to recover its global SFH, as well as the SFH of various regions within the galaxy.

Of course, comparing the "instantaneous" SFR of galaxies observed at various redshift is another way to derive the history of galaxies (provided it can be assumed that the galaxies observed at various redshifts probe the same population). Nowadays, deep surveys have allowed to measure SFRs in many samples up to relatively large redshifts. Interestingly, a SFR-stellar mass relationship seems to exist (with considerable scatter nevertheless) at all redshifts (see, e.g., the compilation in Boissier et al. 2010), providing an important clue on the history of galaxies (more work is however needed to definitively establish its reallity). In a few studies, SFR measurements have been combined with determinations of the molecular gas content of high-redshift galaxies, what allowed to test the validity of the Schmidt Laws described in ➔ Sect. 4.2.4 in the young universe (Bouché et al. 2007; Daddi et al. 2010; Genzel et al. 2010). It was found that a universal Schmidt Law seems to be valid, independently of redshift up to at least z ~ 2.5, including both normal and starburst galaxies. The later ones, however, may require a star formation efficiency larger by a factor 4 at high redshift than for $z \sim 0$. It should be kept in mind that uncertainties stay extremely large (for instance, on the H2-CO conversion factor at high redshift) and the number of studies small.

Finally, it is possible to study the evolution with redshift of the "cosmic" SFR, i.e., the SFR per unit volume averaged over very large scales, to obtain a SFR representative of the whole population of galaxies. The famous "Madau plot" shows such a cosmic SFR density as a function of redshift (Madau et al. 1996). Since then, many works have attempted to add points to the diagram and to interpret it. A compilation of measurements of the cosmic SFR density can be found in Hopkins and Beacom (2006), but new observations are added every year from various sources (e.g., galaxies deep surveys, gamma ray bursts). This "cosmic SFR" is to be related to all the other "cosmic" variables that can be obtained trough observations. Wilkins et al. (2008b) compiled measurements of the cosmic stellar mass density as a function of redshift and studied the coherence of the cosmic SFR and stellar mass densities. A related subject is the amount of metals formed by all the generation of stars: an evolution of the "cosmic metallicity" should be found. Metallicity of high-redshift galaxies are measured in QSO or GRB absorbing systems, but are often difficult to relate to the average metallicity on very large scales. Savaglio et al. (2009) show the evolution with reshift of abundances observed in large column density absorbers (for which precise measurements are achievable), but it is unsure that they are representative of the whole population of galaxies. At lower densities, it is possible to place limits on the cosmic metallicity from ionic abundances (e.g., Songaila 2001). Finally, it is possible to link the density of heavy element to the background radiation emitted by the stars responsible for their formation (e.g., Longair 1995, and references therein).

In fact, the cosmic SFR density (and then the related cosmic stellar mass density and cosmic metallicity) is not linked solely to the actual transformation of gas into stars. It is modulated by the SFR distribution and influenced by the whole range of physical processes affecting galaxy evolution (accretion, interactions, feedback, etc.), discussed in more details in the dedicated chapter.

Acknowledgments

It is a pleasure to thank many collaborators with whom I had the chance to discuss various aspects of star formation during the past 10 years, especially Barry Madore and Nikos Prantos. In the process of writing this chapter, I also had stimulating discussions with my colleagues at the Laboratoire d'Astrophysique de Marseille, especially Alessandro Boselli, Véronique Buat, Denis Burgarella, Sébastien Heinis, Olivier Ilbert, and Olga Cucciati. I am thankful for even the shortest discussions we had, and their encouragements. I also thank Jarle Brinchmann who replied to my e-mails instantaneously with clear answers to my questions.

References

Abramova, O. V., & Zasov, A. V. 2008, Astron. Rep., 52, 257
Argence, B., & Lamareille, F. 2009, A&A, 495, 759
Barnes, J. E. 2004, MNRAS, 350, 798
Bastian, N., Covey, K. R., & Meyer, M. R. 2010, ARA&A, 48, 339
Bell, E. F. 2003, ApJ, 586, 794
Bell, E. F., et al. 2005, ApJ, 625, 23
Bicker, J., & Fritze-v. Alvensleben, U. 2005, A&A, 443, L19
Bigiel, F., Leroy, A., Walter, F., Brinks, E., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2846
Binney, J., & Tremaine, S. 1987, Galactic Dynamics (Princeton: Princeton University Press)
Blitz, L., & Rosolowsky, E. 2006, ApJ, 650, 933
Boissier, S., & Prantzos, N. 1999, MNRAS, 307, 857
Boissier, S., Prantzos, N., Boselli, A., & Gavazzi, G. 2003, MNRAS, 346, 1215
Boissier, S., et al. 2007, ApJS, 173, 524
Boissier, S., et al. 2008, ApJ, 681, 244
Boissier, S., Buat, V., & Ilbert, O. 2010, A&A, 522, A18+
Boquien, M., Duc, P., Braine, J., Brinks, E., Lisenfeld, U., & Charmandaris, V. 2007, A&A, 467, 93
Boquien, M., et al. 2010a, A&A, 518, L70+
Boquien, M., et al. 2010b, ApJ, 713, 626
Boselli, A., Gavazzi, G., Donas, J., & Scodeggio, M. 2001, AJ, 121, 753
Boselli, A., Boissier, S., Cortese, L., Buat, V., Hughes, T. M., & Gavazzi, G. 2009, ApJ, 706, 1527
Bouché, N., et al. 2007, ApJ, 671, 303
Brinchmann, J., Charlot, S., White, S. D. M., Tremonti, C., Kauffmann, G., Heckman, T., & Brinkmann, J. 2004, MNRAS, 351, 1151
Bronfman, L., Casassus, S., May, J., & Nyman, L. 2000, A&A, 358, 521
Buat, V., & Xu, C. 1996, A&A, 306, 61
Buat, V., et al. 2008, A&A, 483, 107
Buat, V., et al. 2010, MNRAS, 409, L1
Bush, S. J., Cox, T. J., Hernquist, L., Thilker, D., & Younger, J. D. 2008, ApJL, 683, L13
Calzetti, D. 1997, The ultraviolet universe at low and high redshift, in AIP Conf. Ser. 408, ed. W. H. Waller (New York: AIP), 403–412
Calzetti, D., & Kennicutt, R. C. 2009, PASP, 121, 937
Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582
Calzetti, D., et al. 2007, ApJ, 666, 870
Calzetti, D., Sheth, K., Churchwell, E., & Jackson, J. 2009, in The Evolving ISM in the Milky Way and Nearby Galaxies, Chicago
Calzetti, D., et al. 2010, ApJ, 714, 1256
Cassata, P., et al. 2011, A&A, 525, A143+
Chabrier, G., 2003, PASP, 115, 763
Charlot, S., & Fall, S. M. 2000, ApJ, 539, 718
Charlot, S., & Longhetti, M. 2001, MNRAS, 323, 887
Charlot, S., Kauffmann, G., Longhetti, M., Tresse, L., White, S. D. M., Maddox, S. J., & Fall, S. M. 2002, MNRAS, 330, 876
Cignoni, M., Degl'Innocenti, S., Prada Moroni, P. G., & Shore, S. N. 2006, A&A, 459, 783

Condon, J. J. 1992, ARA&A, 30, 575

Corbelli, E. 2003, MNRAS, 342, 199

Cortese, L., Boselli, A., Franzetti, P., Decarli, R., Gavazzi, G., Boissier, S., & Buat, V. 2008, MNRAS, 386, 1157

Cowie, L. L. 1981, ApJ, 245, 66

Cucciati, O., et al. 2012, A&A, 539, 31

Daddi, E., et al. 2010, ApJL, 714, L118

Deharveng, J., et al. 2008, ApJ, 680, 1072

Di Matteo, P., Bournaud, F., Martig, M., Combes, F., Melchior, A., & Semelin, B. 2008, A&A, 492, 31

Dopita, M. A., & Ryder, S. D. 1994, ApJ, 430, 163

Elmegreen, B. G. 1979, ApJ, 231, 372

Elmegreen, B. G. 1993a, in Star Formation, Galaxies and the Interstellar Medium, ed. J. Franco, F. Ferrini, & G. Tenorio-Tagle (Cambridge/New York: Cambridge University Press), 337–348

Elmegreen, B. G. 1993b, ApJ, 411, 170

Elmegreen, B. G., & Hunter, D. A. 2006, ApJ, 636, 712

Feldmann, R., Gnedin, N. Y., & Kravtsov, A. V. 2011, ApJ, 732, 115

Ferguson, A. M. N., Wyse, R. F. G., Gallagher, J. S., & Hunter, D. A. 1998, ApJL, 506, L19

Fuchs, B., Jahreiß, H., & Flynn, C. 2009, AJ, 137, 266

Gallagher, J. S., Bushouse, H., & Hunter, D. A. 1989, AJ, 97, 700

Gao, Y., & Solomon, P. M. 2004, ApJ, 606, 271

Genzel, R., et al. 2010, MNRAS, 407, 2091

Giavalisco, M., Koratkar, A., & Calzetti, D. 1996, ApJ, 466, 831

Gil de Paz, A., et al. 2005, ApJL, 627, L29

Gilbank, D. G., Baldry, I. K., Balogh, M. L., Glazebrook, K., & Bower, R. G. 2010, MNRAS, 405, 2594

Goddard, Q. E., Kennicutt, R. C., & Ryan-Weber, E. V. 2010, MNRAS, 405, 2791

Grimm, H., Gilfanov, M., & Sunyaev, R. 2003, MNRAS, 339, 793

Gronwall, C., et al. 2007, ApJ, 667, 79

Guibert, J., Lequeux, J., & Viallefond, F. 1978, A&A, 68, 1

Harris, J., & Zaritsky, D. 2009, AJ, 138, 1243

Hernandez, X., Valls-Gabaud, D., & Gilmore, G. 2000, MNRAS, 316, 605

Hirashita, H., Buat, V., & Inoue, A. K. 2003, A&A, 410, 83

Hopkins, A. M., & Beacom, J. F. 2006, ApJ, 651, 142

Hopkins, A. M., Connolly, A. J., Haarsma, D. B., & Cram, L. E. 2001, AJ, 122, 288

Hunter, D. A., Elmegreen, B. G., & Baker, A. L. 1998, ApJ, 493, 595

Iglesias-Páramo, J., Boselli, A., Gavazzi, G., & Zaccardo, A. 2004, A&A, 421, 887

Iglesias-Páramo, J., et al. 2006, ApJS, 164, 38

James, P. A., et al. 2004, A&A, 414, 23

Jog, C. J., & Solomon, P. M. 1984, ApJ, 276, 114

Karachentsev, I. D., & Kaisin, S. S. 2010, AJ, 140, 1241

Kennicutt, R. C., Jr. 1989, ApJ, 344, 685

Kennicutt, R. C. 1997, The interstellar medium in galaxies, in Astrophysics and Space Science Library, Vol. 161 (Dordrecht: Astrophysics and Space Science Library), 171–195

Kennicutt, R. C., Jr. 1998a, ARA&A, 36, 189

Kennicutt, R. C., Jr. 1998b, ApJ, 498, 541

Kennicutt, R. C., Jr. et al. 2007, ApJ, 671, 333

Kennicutt, R. C., Jr., Lee, J. C., Funes, José G., S. J., Sakai, S., & Akiyama, S. 2008, ApJS, 178, 247

Kennicutt, R. C., et al. 2009, ApJ, 703, 1672

Kewley, L. J., Geller, M. J., & Jansen, R. A. 2004, AJ, 127, 2002

Koda, J. 2008, Formation and evolution of galaxy disks, in ASP Conf. Ser. 396, ed. J. G. Funes, & E. M. Corsini (San Francisco: Astronomical Society of the Pacific), 97–+

Kong, X., Charlot, S., Brinchmann, J., & Fall, S. M. 2004, MNRAS, 349, 769

Kroupa, P. 2001, MNRAS, 322, 231

Krumholz, M. R., & McKee, C. F. 2008, Nature, 451, 1082

Krumholz, M. R., McKee, C. F., & Tumlinson, J. 2009, ApJ, 699, 850

Larson, R. 1992, in Star Formation in Stellar Systems, ed. G. Tenorio-Tagle, M. Prieto, & F. Sanchez (Cambridge/New York: Cambridge University Press), 125–+

Lee, J. C., et al. 2009, ApJ, 706, 599

Leroy, A. K., Walter, F., Brinks, E., Bigiel, F., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2782

Lisenfeld, U., Braine, J., Duc, P., Charmandaris, V., Vallejo, O., Leon, S., & Brinks, E. 2001, in Dwarf Galaxies and Their Environment, ed. K. S. de Boer, R.-J. Dettmar, & U. Klein (Aachen: Shaker), 273–+

Liu, G., et al. 2011, ApJ, 735, 63

Longair, M. S. 1995, in Extragalactic Background Radiation Meeting, ed. D. Calzetti, M. Livio, & P. Madau (Cambridge: Cambridge University Press), 223–236

Luna, A., Bronfman, L., Carrasco, L., & May, J. 2006, ApJ, 641, 938

Madau, P., Ferguson, H. C., Dickinson, M. E., Giavalisco, M., Steidel, C. C., & Fruchter, A. 1996, MNRAS, 283, 1388

Madore, B. F. 1977, MNRAS, 178, 1

Madore, B. F. 2010, ApJL, 716, L131

Madore, B. F., van den Bergh, S., & Rogstad, D. H. 1974, ApJ, 191, 317

Malhotra, S., et al. 2001, ApJ, 561, 766

Martin, C. L., & Kennicutt, R. C., Jr. 2001, ApJ, 555, 301

McKee, C. F., & Ostriker, E. C. 2007, ARA&A, 45, 565
McQuinn, K. B. W., et al. 2010, ApJ, 724, 49
Meurer, G. R., Heckman, T. M., Leitherer, C., Kinney, A., Robert, C., & Garnett, D. R. 1995, AJ, 110, 2665
Meurer, G. R., Heckman, T. M., & Calzetti, D. 1999, ApJ, 521, 64
Mo, M., van den Bosch, F., & White, S. 2009, Galaxies Formation and Evolution (Cambridge: Cambridge University Press)
Monaco, P., Murante, G., Bornagi, S., & Dolag, K. 2012, MNRAS, 421, 2485
Mouhcine, M., Lewis, I., Jones, B., Lamareille, F., Maddox, S. J., & Contini, T. 2005, MNRAS, 362, 1143
Muñoz-Mateos, J. C., Gil de Paz, A., Boissier, S., Zamorano, J., Jarrett, T., Gallego, J., & Madore, B. F. 2007, ApJ, 658, 1006
Muñoz-Mateos, J. C., et al. 2009, ApJ, 701, 1965
Murgia, M., Crapsi, A., Moscadelli, L., & Gregorini, L. 2002, A&A, 385, 412
Noeske, K. G., et al. 2007, ApJL, 660, L47
Osterbrock, D. E., & Ferland, G. J. 2006, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei, ed. D. E. Osterbrock, & G. J. Ferland (Sausalito: University Science Books)
Péroux, C., Bouché, N., Kulkarni, V. P., York, D. G., & Vladilo, G. 2011, MNRAS, 410, 2251
Pflamm-Altenburg, J., Weidner, C., & Kroupa, P. 2007, ApJ, 671, 1550
Quirk, W. J. 1972, ApJL, 176, L9
Ranalli, P., Comastri, A., & Setti, G. 2003, A&A, 399, 39
Robitaille, T. P., & Whitney, B. A. 2010, ApJL, 710, L11
Rocha-Pinto, H. J., Scalo, J., Maciel, W. J., & Flynn, C. 2000, A&A, 358, 869
Rodriguez-Fernandez, N. J., Braine, J., Brouillet, N., & Combes, F. 2006, A&A, 453, 77
Salim, S., et al. 2005, ApJL, 619, L39
Salpeter, E. E. 1955, ApJ, 121, 161
Sandage, A. 1986a, A&A, 161, 89
Sandage, A. 1986b, A&A, 181, 89
Sanduleak, N. 1969, AJ, 74, 47
Savaglio, S., Glazebrook, K., & Le Borgne, D. 2009, ApJ, 691, 182
Schaye, J. 2004, ApJ, 609, 667
Schiminovich, D., et al. 2010, MNRAS, 408, 919
Schmidt, M. 1959, ApJ, 129, 243
Searle, L., Sargent, W. L. W., & Bagnuolo, W. G. 1973, ApJ, 179, 427
Seibert, M., et al. 2005, ApJL, 619, L55
Seigar, M. S. 2005, MNRAS, 361, L20
Shi, Y., et al. 2011, ApJ, 733, 87
Songaila, A. 2001, ApJL, 561, L153
Tamburro, D., Rix, H., Walter, F., Brinks, E., de Blok, W. J. G., Kennicutt, R. C., & Mac Low, M. 2008, AJ, 136, 2872
Tan, J. C. 2000, ApJ, 536, 173
Teyssier, R., Chapon, D., & Bournaud, F. 2010, ApJL, 720, L149
Thilker, D. A., et al. 2005, ApJL, 619, L79
Thilker, D. A., et al. 2007a, ApJS, 173, 538
Thilker, D. A., et al. 2007b, ApJS, 173, 572
Tolstoy, E., Hill, V., & Tosi, M. 2009, ARA&A, 47, 371
Toomre, A. 1964, ApJ, 139, 1217
van den Bergh, S. 1999, A&AR, 9, 273
Verley, S., Corbelli, E., Giovanardi, C., & Hunt, L. K. 2010a, A&A, 510, A64+
Verley, S., et al. 2010b, A&A, 518, L68+
Wang, B., & Silk, J. 1994, ApJ, 427, 759
Weilbacher, P. M., & Fritze-v. Alvensleben, U. 2001, A&A, 373, L9
Whitney, B. A., et al. 2008, AJ, 136, 18
Wilkins, S. M., Hopkins, A. M., Trentham, N., & Tojeiro, R. 2008a, MNRAS, 391, 363
Wilkins, S. M., Trentham, N., & Hopkins, A. M. 2008b, MNRAS, 385, 687
Wong, T., & Blitz, L. 2002, ApJ, 569, 157
Wyder, T. K., et al. 2009, ApJ, 696, 1834
Wyse, R. F. G. 1986, ApJL, 311, L41
Wyse, R. F. G. 2009, in IAU Symp., Vol. 258, ed. E. E. Mamajek, D. R. Soderblom, & R. F. G. Wyse, 11–22
Wyse, R. F. G., & Silk, J. 1989, ApJ, 339, 700

4 The Cool ISM in Galaxies

Jan M. van der Hulst[1] · *W. J. G. de Blok*[2]
[1]Radio Astronomy, Kapteyn Astronomical Institute, University of Groningen, Groningen, The Netherlands
[2]Astronomy Group, ASTRON, Netherlands Foundation for Radio Astronomy, Dwingeloo, The Netherlands

1 Introduction 184

2 The Neutral Hydrogen (*H I*) in Galaxies 185
2.1 The H I Physics and Observables 185
2.2 The Distribution of H I in Galaxies 188
2.3 The Warm and Cold ISM in Galaxies 192

3 Star Formation and the ISM 193

4 Accretion, Feedback and the Environment 199

References 202

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_4, © Springer Science+Business Media Dordrecht 2013

Abstract: This chapter describes the different constituents of the observable interstellar medium (ISM) in galaxies and reviews the relationships between the ISM and the star formation in galaxies. The emphasis is on the component which is most widespread and most easily observable, the neutral atomic hydrogen (H I). It briefly touches upon effects of the environment and of the interplay between star formation and the ISM (feedback).

1 Introduction

The best census of the different components of the interstellar medium (ISM) exists for the Milky Way, the result of close to half a century of observational and theoretical studies. Much of this is well described in various chapters of the Handbook (➜ Chap. 10 of Volume 5). The ISM components range from various forms of the most abundant specious, neutral hydrogen (H I), ionized hydrogen (H II), and molecular hydrogen (H_2), and a large variety of complex molecules and dust grains, to the more energetic components of the cosmic ray population. In this chapter, we restrict ourselves to the neutral hydrogen component of the ISM in galaxies and only briefly address some of the other components such as the ionized and molecular hydrogen, and the dust.

The ISM plays a crucial role in the process of star formation throughout a galaxy's lifetime. How efficiently star formation proceeds depends on the local properties of the ISM. These are in turn affected by the star formation which provides feedback via stellar winds and supernova explosions. This feedback is an important regulating mechanism for the ongoing process of the buildup of galaxies throughout cosmic time. Detailed imaging of the ISM, in particular in H I is now available for many objects in the nearby universe. This chapter will therefore briefly review the current ideas and observational evidence for the physical connection between the local properties of the ISM and the star formation.

Although such constituents of the ISM as dust and ionized gas have long been known in the Galaxy, it was not until the detection of the first 21-cm line emission of H I (Ewen and Purcell 1951; Muller and Oort 1951; Pawsey 1951), following the prediction by van de Hulst (1945), opened a full perspective on the structure, kinematics, and physics of the H I in the Galaxy (see ➜ Chap. 11 of Volume 5). The first pioneering observations of H I in other galaxies (Raimond and Volders 1957; Volders 1959; Volders and Högbom 1961) were the beginning of several decades of galaxy surveys in the H I line using single-dish radio telescopes of increasing diameter. This led to insight into the global properties of H I in galaxies, well summarized in Roberts and Haynes (1994) following the much earlier reviews of Roberts (1963, 1975).

The advent of earth-rotation synthesis radio telescopes drastically changed the field, as it became possible to image the H I in galaxies with angular resolutions of a few arcminutes originally (Baldwin et al. 1971; Rogstad and Shostak 1971; Wright et al. 1972) with quality improving over the years to ~5″ (Walter et al. 2008). By 2012 some 500 galaxies have been imaged in the H I line, some individually, some as part of large observing programs such as WHISP (van der Hulst et al. 2001; García-Ruiz et al. 2002; Swaters et al. 2002; Noordermeer et al. 2005) and THINGS (Walter et al. 2008). Although a main focus of the earlier work was on using the H I kinematics to determine the mass distributions of galaxies, in particular the dark matter content (Sanders 2010), this chapter will primarily discuss the distribution of the H I in relation to other components of the ISM, and in particular in relation to the star formation in galaxies, using the latest results from surveys such as WHISP and THINGS in H I, and BIMA-SONG (Helfer et al. 2003) and HERACLES (Leroy et al. 2008, 2009) in CO.

This chapter will not repeat the information of many previous reviews, nor will it review recent low-resolution surveys of H I in galaxies such as HIPASS (Meyer et al. 2004; Zwaan et al. 2004; Koribalski et al. 2004) or ALFALFA (Giovanelli et al. 2005; Toribio et al. 2011). It will concentrate instead on the resolved distribution of the ISM as traced by the H I, as there now is a wealth of information from high-resolution H I imaging of a few hundred galaxies. Furthermore, it will discuss the relation of the H I and other ISM components to the star formation in galaxies and touch upon the role of the star formation and the environment in shaping the distribution of the H I. This chapter provides an overview of the current knowledge of the cool ISM and its relation to star formation in galaxies, but does not pretend to be complete.

2 The Neutral Hydrogen (H I) in Galaxies

This section first provides a brief discussion of the relevant physics of the H I and associated observables, followed by an overview of the detailed distribution of H I in galaxies, and a brief description of the warm and cold components of the ISM.

2.1 The H I Physics and Observables

The basic physics of the H I atom is well described in ➲ Chap. 11 of Volume 5 and in Walterbos and Braun (1996). A full treatment of the level population of the hydrogen atom is given in Field (1959). In the ISM, collisions dominate the excitation of the fine-structure 21-cm H I line. Resolved observations of H I in galaxies provide a set of images of the H I at a series of frequencies (i.e., Doppler velocities) determined by the spectral resolution and observing frequency used. Each image basically provides the brightness distribution of the H I line at a particular Doppler velocity. For each line of sight then an H I profile can be constructed which can be integrated over all velocities to determine a column density. The realization that the H I in galaxy disks is predominantly optically thin allows us to determine H I column densities directly from the measured brightness temperatures of the H I emission using the following equation:

$$\frac{N_{\mathrm{H\,I}}}{\mathrm{cm}^{-2}} = 1.823 \times 10^{18} \int \frac{T_B(v)}{\mathrm{K\,kms}^{-1}} dv, \tag{4.1}$$

where the brightness temperature $T_B(v)$ is related to the flux per beam ($S(v)$, the usual units used in synthesis maps) and the beam area (Ω_B) via

$$T_B(v) = 606(\lambda/21.1\mathrm{cm})^2 \frac{\mathrm{S(v)\ (mJy)}}{\Omega_{\mathrm{B}}\ \mathrm{(sq.arcsec)}} \tag{4.2}$$

The column density basically is the integral over velocity of all emission along a given line of sight within one resolution element. Commonly, it is determined by calculating the 0th moment of the measured H I profiles. Higher order moments can also be calculated and are a measure of a characteristic radial velocity (1st moment) and velocity dispersion (2nd moment). These higher order moments are not necessarily the best measures of velocity and velocity dispersion. Better results are obtained by fitting profiles with a single or multiple Gaussians or Gauss-Hermite polynomials (Noordermeer et al. 2005; Swaters et al. 2002).

In addition to determining observational parameters of each line of sight, one can also integrate the H I signal spatially to determine the total flux in each channel for a given object (e.g., an entire galaxy, a companion feature, or a feature within a galaxy disk). These fluxes can be integrated over velocity to determine the total mass, which is related to the total flux integral as

$$\frac{M_{\mathrm{H\,I}}}{M_\odot} = 236\left(\frac{D}{\mathrm{Mpc}}\right)^2 \int \frac{S(v)}{\mathrm{mJy}}\,\frac{dv}{\mathrm{kms}^{-1}} \tag{4.3}$$

When calculating detection limits based on the rms noise of the observation, one has to assume a profile velocity width as both $N_{\mathrm{H\,I}}$ and $M_{\mathrm{H\,I}}$ are integrated over velocity. This has led to a range of limit definitions in the literature as some authors integrate over the velocity resolution while others use (different) estimates of the expected profile width for the objects under study. It would be better to introduce uniformity and quote limits in $N_{\mathrm{H\,I}}/\Delta v$ and $M_{\mathrm{H\,I}}/\Delta v$ rather than $N_{\mathrm{H\,I}}$ and $M_{\mathrm{H\,I}}$ to avoid this practical problem.

For H I absorption, the situation is more complex as H I absorption measurements basically probe optical depth:

$$\tau(v) = 5.49 \times 10^{-14}\frac{N_{\mathrm{H\,I}}}{T_s}P(v) \tag{4.4}$$

with a Maxwellian velocity distribution $P(v)$ given by

$$P(v) = \frac{1}{\sqrt{\pi}.b}\,e^{(-v/b)^2} \tag{4.5}$$

with

$$b = \sqrt{\frac{2kT_k}{m_{\mathrm{H\,I}}}}, \tag{4.6}$$

where $\tau(v)$ is the optical depth and T_s and T_k are the spin temperature and the kinetic temperature of the H I, respectively. Because the optical depth depends on both the column density ($N_{\mathrm{H\,I}}$) and the spin temperature, it is not possible to interpret absorption line measurements unambiguously. This is a major limitation which can only be resolved if one has an independent measurement of HI emission along the same line of sight. For small optical depths, the absorption and emission observations can be combined to solve for $N_{\mathrm{H\,I}}$ and T_s independently. Such observations do not generally exist, however, so an assumption about the spin temperature is required to estimate column densities from absorption line measurements.

An example of the H I distribution of a spiral galaxy is given in *Fig. 4-1*. It shows NGC 5055 in H I, visible light, and UV (GALEX). The H I reveals a wealth of structure, not only in the inner part of the galaxy, coincident with the bright optical disk, but also in the outer parts where faint light is present and structures in UV light indicate the presence of star formation. This H I observation is rather deep. Most H I synthesis observations in the literature do not probe much deeper than column densities of $N_{\mathrm{H\,I}} \sim 10^{20}\mathrm{cm}^{-2}$ for the typical sensitivities and angular resolutions used. Synthesis observations offer the flexibility of improving the column density sensitivity by changing the weights of the u, v data. Giving lower weight to the longer baselines lowers the resolution (increases Ω_B) and enhances the surface brightness sensitivity and hence column density sensitivity. This is illustrated further in *Fig. 4-2* which shows the H I distributions of NGC 6946 at two resolutions: 6″ and 60″. It is clear that the 60″ image brings out fainter features at the expense of resolution, especially in the outer parts. Most notable is the faint tail in the northwest which is interpreted as the remnant of an infalling gas complex (Boomsma et al. 2008). These images result from long integrations at the Very Large Array (VLA) and Westerbork Synthesis Radio Telescope (WSRT).

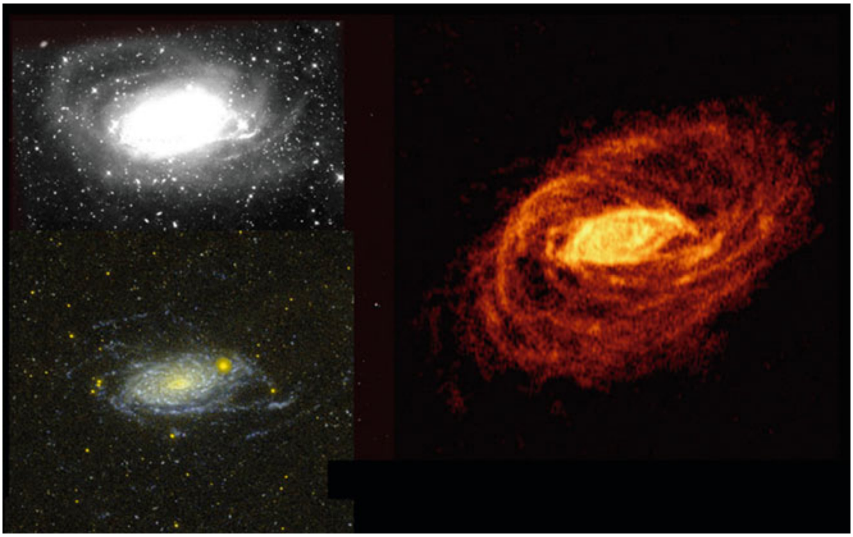


Fig. 4-1
Deep optical (Martínez-Delgado et al. 2010), GALEX UV (Thilker et al. 2007), and WSRT H I (Battaglia et al. 2005) images of the warped galaxy NGC 5055. The H I intensities range from a limiting column density of about 3 × 10^{19} cm^{-2} to a peak of 1 × 10^{21} cm^{-2} (Note the extent of the H I but also the presence of faint optical features and star formation in the outer parts with its wealth of structure)

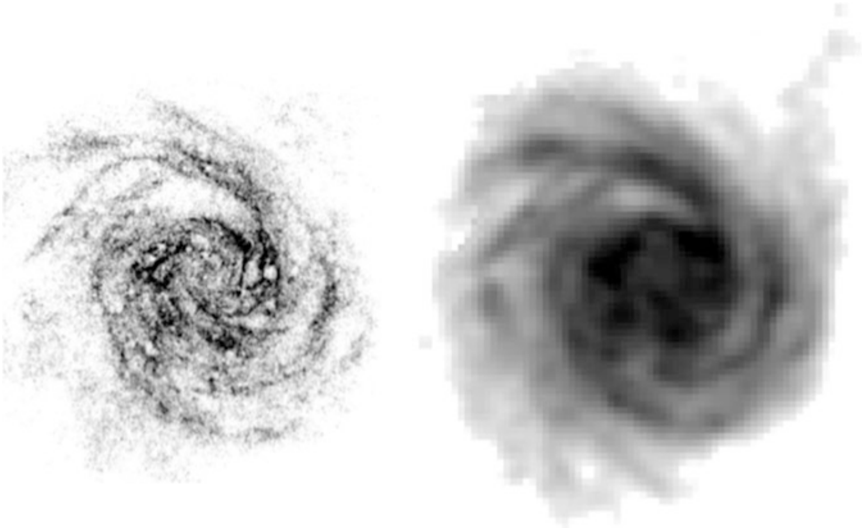


Fig. 4-2
H I images of the galaxy NGC 6946 at two different resolutions. *Left panel*: 6″ resolution VLA data from THINGS (Walter et al. 2008). *Right panel*: 60″ resolution WSRT data from Boomsma et al. (2008) (Note the faint extension in the northwest, possibly a recent accretion event)

2.2 The Distribution of H I in Galaxies

Observations of H I in other galaxies have clearly confirmed the picture that emerged from studies of the Milky Way: most of the H I is in a thin (~ few hundred pc scale height), rotating disk, typically more extended radially than the observable distribution of star light. This disk may flare and warp in the outer parts as does the H I disk of the Milky Way (Levine et al. 2006; Kerr 1969, and references therein). In addition, evidence is mounting for the presence of extraplanar H I: either gas thrown out of the disk by star formation activity (the "Galactic Fountain," Bregman 1980), or gas accreted from satellite galaxies (such as the Magellanic Stream around the Milky Way, Putman et al. 2003) or from the cosmic web (Sancisi et al. 2008). Edge-on galaxies provide the best view of both the thin H I disk and such extraplanar H I. Prominent, well-studied examples are NGC 891 and NGC 4565 (Oosterloo et al. 2007a; Rupen 1991). In NGC 891, observations of increasing depth provided the first evidence for extraplanar H I. Part of this H I is in a more slowly rotating, thick disk reaching scale heights of a few kpc; another part consists of more patchy filaments, reaching heights of 10 kpc above the plane, and is kinematically less well behaved. These components are illustrated in ➋ *Fig. 4-3* and will be discussed in more detail later.

The overall radial surface density (i.e., deprojected column density) distribution of H I in the disks of spiral galaxies is rather flat within the optical disk, with an approximately exponential decline in the outer parts (Swaters et al. 2002; Bigiel et al. 2008). Early-type spiral galaxies tend to have a deficiency of H I surface density in the inner parts, often compensated by the presence of large amounts of H_2 (Noordermeer et al. 2005). Typical average surface densities

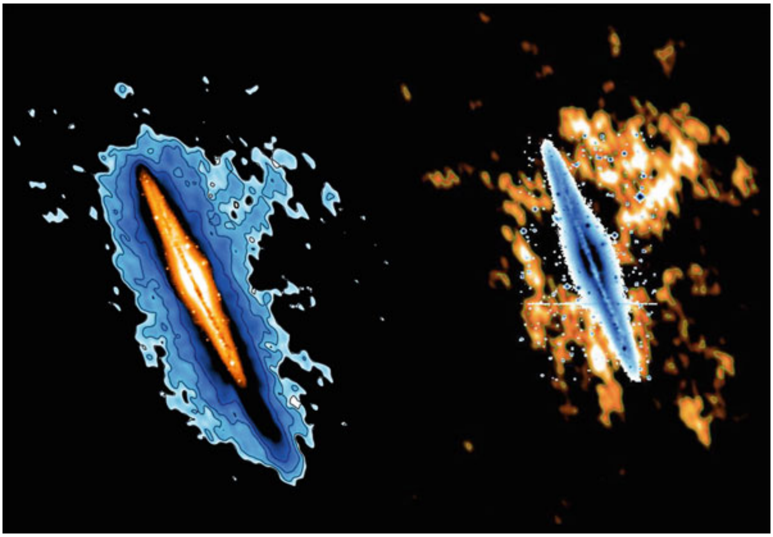


Fig. 4-3

Deep WSRT H I images of NGC 891 (Oosterloo et al. 2007a). *Right panel* shows the integrated H I image superposed on the optical image of the galaxy. At this resolution and sensitivity, the thick, lopsided H I distribution is very clear. The *right panel* shows only the H I that is not corotating with the disk. This anomalous-velocity gas is everywhere, but most prominent and extended in the northwest. Its H I mass is ~10^8 $M_\odot$, a few percent of the total H I mass of the galaxy

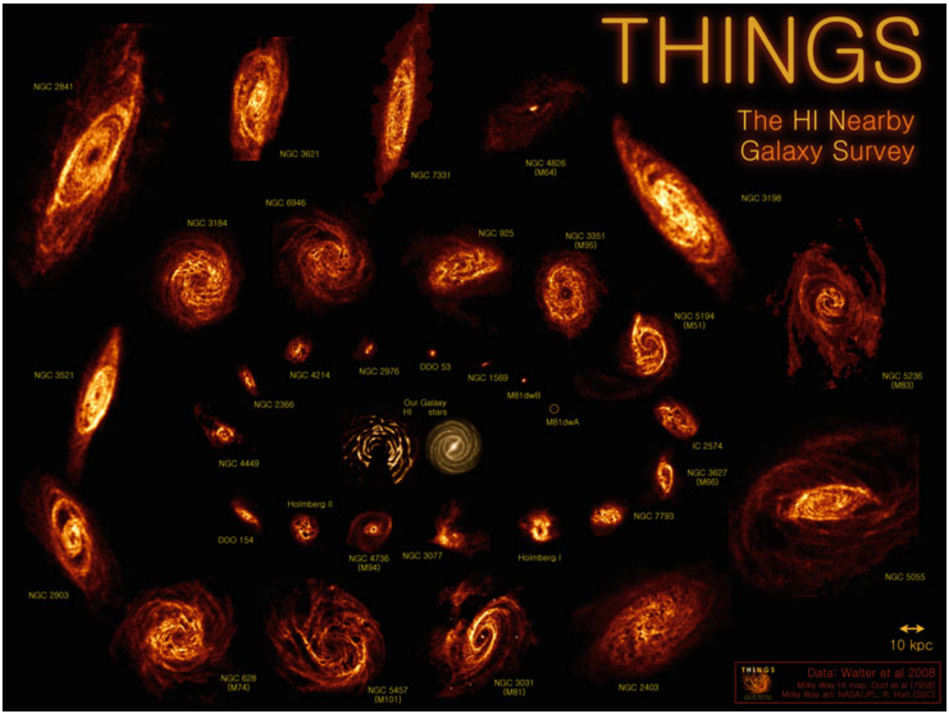


Fig. 4-4
H I images of 34 galaxies observed in the THINGS program (Walter et al. 2008). All galaxies are shown on the same linear scale, indicated by the *arrow* in the *lower right corner*. For comparison, the H I disk of the Milky Way is displayed on the same scale in the center of the figure

in the disk are a few $M_\odot\ \mathrm{pc}^{-2}$ in the inner parts, with local maxima of over $10\ M_\odot\ \mathrm{pc}^{-2}$. Blitz and Rosolowsky (2006) investigated the effect of pressure in the ISM and suggested that above $10\ M_\odot\ \mathrm{pc}^{-2}$ most of the hydrogen turns into molecular form. This is actually confirmed by observations (Bigiel et al. 2008).

H I disks exhibit a wealth of structure. This is very obvious in a bird's-eye view of 34 different galaxies from the THINGS survey (➋ *Fig. 4-4*). In the inner parts, they exhibit the same structure as seen optically, traced by the recent star formation (spiral arms, spiral arm segments, filaments, etc.). The outer H I disks also exhibit structure, often with spiral arm-like features similar to the inner parts. Also in the outer disk, the brightest H I structures coincide often with locations of star formation as shown by the UV emission detected by GALEX in the outer disks of galaxies (Thilker et al. 2005, 2007). This is very clear in ➋ *Figs. 4-1* and ➋ *4-9*.

When examined globally, many galaxies exhibit asymmetries, both kinematically and spatially. Sancisi et al. (2008) examined some 300 galaxies in the WHISP sample and concluded that half of the objects are asymmetric. A more quantitative analysis by van Eymeren et al. (2011a, b) shows that more than 60% of the WHISP galaxies show kinematic and/or morphological asymmetries. The latter studies show that kinematic lopsidedness is more common than the occurrence of morphological asymmetries. The cause for asymmetries is not well established. Sancisi et al. (2008) suggest that accretion or minor mergers are the main cause. This could be fallback of material in tidally interacting systems, though not in every case as not all lopsided galaxies are interacting systems. van Eymeren et al. (2011b) investigate the dependence of the degree of lopsidedness on the environment and confirm that there is no clear effect. So if tidal interactions are the only way to induce lopsidedness, it needs to be long-lived.

In addition to global asymmetries, a large fraction of H I disks of galaxies appear warped (García-Ruiz et al. 2002; Sancisi et al. 2008). Warped H I disks are best studied in edge-on galaxies but can also be found and quantified through their signature in the H I velocity fields as described by Briggs (1990) and more recently by Józsa (2007) and Józsa et al. (2007). The warps usually start at the edge of the optical disk (García-Ruiz et al. 2002). As with the asymmetries in general, warps need to be long-lived if tidal interactions are the only cause, suggesting other mechanisms such as accretion and minor mergers (Sancisi et al. 2008). The warped H I layer of the Milky Way increases in thickness toward the outer parts (Levine et al. 2006; Kalberla and Kerp 2009). This thickening has not been established as well for other galaxies. Some well-studied edge-on galaxies have been shown to exhibit flaring H I layers (Sancisi and Allen 1979; Olling 1996; O'Brien et al. 2010), indicating that flaring H I disks may be very common. Further investigations combining accurate measurement of both the velocity dispersion of the gas and the thickness of the H I layer are required to confirm this trend for a larger number of galaxies. For a more detailed overview of the methods and results see Kregel and van der Kruit (2005 and references therein), O'Brien et al. (2010), and the review by van der Kruit and Freeman (2011). There is an indication that low-luminosity dwarf galaxies have thicker H I disks than normally found in spiral galaxies (Roychowdhury et al. 2009, 2010).

Detailed studies at increasing sensitivity and resolution of the nearby edge-on galaxy NGC 891 (Sancisi and Allen 1979; Swaters et al. 1997; Oosterloo et al. 2007a) have shown that there exist a thick H I disk and an H I halo. They are corotating with the disk, though the rotation slows down with increasing height above the plane. This decrease is too large (20–40 km s^{-1}) to be explained by asymmetric drift. The existence of a thick disk/halo can also be detected in moderately inclined galaxies as it becomes manifest by its slower rotation as gas at lower velocities in position-velocity diagrams (termed the "beard" by Schaap et al. 2000, who first identified and studied this component in NGC 2403). Fraternali et al. (2001, 2002) made a detailed study of this phenomenon in NGC 2403 using more sensitive H I data. ➋ *Figure 4-5* shows this "beard" in NGC 2403. Modeling showed that also in this galaxy the thick disk is rotating more slowly by a few tens of km s^{-1} as is the case for NGC 891.

The indication for the possible existence of an extended gaseous halo around galaxies originally came from the detection of galactic absorption lines in the direction of bright stars in the Magellanic Clouds that could be ascribed to hot gas in a galactic "corona" (Savage and de Boer 1979). This component, often termed the warm ionized medium (WIM) or the diffuse ionized gas (DIG), has also been shown to be present in the disks and halos of other galaxies (Dettmar 1990; Rand et al. 1990). This thick WIM component in galaxies is thought to be fed by stellar winds and supernova explosions, though the number of "chimney"-like structures expected based on models such as those of Norman and Ikeuchi (1989) has not yet been confirmed by H$_\alpha$ imaging of edge-on galaxies. This phenomenon is also supposed to drive the outflow of neutral gas and be at the root of the "Galactic Fountain" mechanism (Bregman 1980; Kahn 1994; Rosen and Bregman 1995) in the disk-halo interface.

A related phenomenon discovered along with the thick H I disk and H I halo is the presence of gas complexes outside the thin H I disk. These were noted as either gas moving at velocities different from the general rotation in face-on galaxies (M 101: van der Hulst and Sancisi 1988; NGC 6946: Kamphuis and Sancisi 1993; Boomsma et al. 2008) or as features above the disk in edge-on galaxies (NGC 891: Oosterloo et al. 2007a). This anomalous-velocity gas appears to be

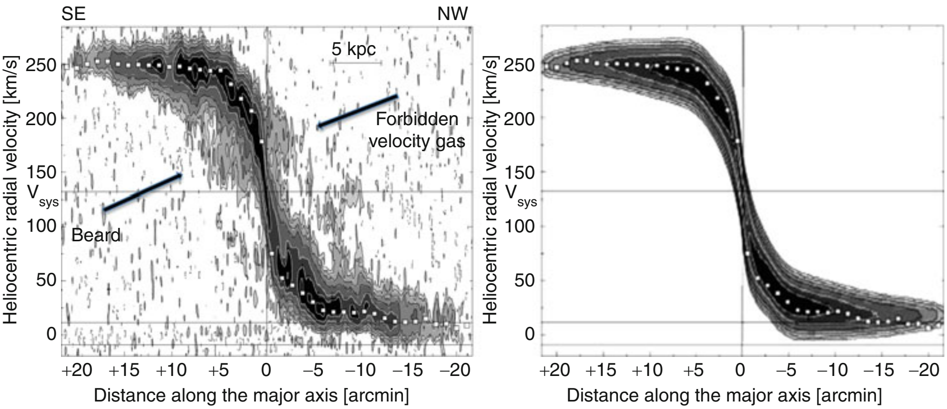


Fig. 4-5

Illustration of the "beard" phenomenon in the galaxy NGC 2403. The *left panel* shows a position-velocity cut along the major axis of NGC 2403 (adopted from Fraternali et al. 2001). Contours are −0.4, 0.4, 1, 2, 4, 10, 20, and 40 mJy beam^{-1}. The *white symbols* represent the projected rotation curve, overplotted on the H I data. The *right panel* shows a model position-velocity diagram for a thin H I disk for comparison. The H I at anomalous velocities is marked by the *arrows* in the *left panel*. The two components are the "beard," moving at velocities slower that the local rotation, and occasional H I structures moving at forbidden velocities

a mix of gas removed from the disk by a "Galactic Fountain" mechanism, related to massive star formation processes in the disk, and steady accretion of intergalactic gas or tidal debris, much like the mechanisms proposed for the presence of the high velocity clouds and the Magellanic Stream around the Milky Way (for more details see the review by Sancisi et al. 2008). The amount of H I in these features typically is of the order of ~10^8 $M_\odot$ or less than 10% of the total amount of H I in a typical disk galaxy.

The discovery of these features required long integrations (~100 h on current synthesis telescopes) and they have thus far been found only in a small number of galaxies. Large surveys of galaxies such as WHISP (van der Hulst et al. 2001) and THINGS (Walter et al. 2008) do not reach the required H I column density sensitivity well below 10^{20} cm^{-2}. A recent deep survey of ~20 galaxies (HALOGAS, Heald et al. 2011) is expected to provide more insight in these phenomena.

Detailed studies of H I in galaxies have been expanded to early-type galaxies (ETGs, usually S0s and ellipticals), systems which have less copious amounts of H I relative to their stellar or total mass. The picture that emerges here (Oosterloo et al. 2007b, c, 2010a; Serra et al. 2012) is that roughly half of the ETGs in the field are detected in H I with detection limits of a few times 10^6 $M_\odot$. About half of these exhibit large, regularly rotating H I disks or rings, while many objects show the presence of tails and outlying H I structures reminiscent of tidal effects and accretion.

In dense environments such as the Virgo cluster, the H I properties of galaxies change drastically as a result of tidal interactions and interaction with the intracluster medium (e.g., see Chung et al. 2009; Solanes et al. 2001). In Virgo, only 10% of the ETGs are detected (Serra et al. 2012), while the spiral galaxies show truncated H I disks and clear signs of stripping and interactions (Chung et al. 2009). These effects will be discussed in more detail in Sect. 4.

At the other extreme, there are the low-luminosity and low-surface-brightness galaxies in the field. Some 65 small, low-luminosity ($M_B > -15$) dwarf irregular galaxies have been studied recently with the GMRT (the Faint Irregular Galaxies GMRT Survey (FIGGS), Begum et al. 2005, 2008) and confirm existing trends in that the H I to optical sizes of galaxies increases toward lower optical luminosities (implying also an increase in M_{HI}/L_B with decreasing luminosity). These low-luminosity systems continue to obey the M_{HI}/D_{HI} relation first demonstrated by Broeils and Rhee (1997), indicating that the average H I surface densities are similar to those in the H I disks of more luminous galaxies. Complementary surveys, carried out with the VLA and focusing on studies of the H I in nearby, low-mass star-forming galaxies, are LITTLE-THINGS (Hunter et al. 2011) and VLA/ANGST (Ott et al. 2008).

The classical low-surface-brightness galaxies (van der Hulst et al. 1993; de Blok et al. 1996) appear to fill the gap between these very-low-luminosity systems and the galaxies of normal surface brightness and normal luminosity in terms of their properties.

2.3 The Warm and Cold ISM in Galaxies

The presence of ionized hydrogen around OB stars, in the form of H II regions, has been known for a long time. The idea that in addition the ISM in the Galaxy may have a more widespread, hot, ionized hydrogen component came from the interpretation of the low-frequency turnover of the spectrum of the Galactic synchrotron emission as free-free absorption (Hoyle and Ellis 1963). Ten years later, the presence of the diffuse ionized interstellar medium was clearly demonstrated from the detection of ubiquitous Hα emission (Reynolds 1971; Reynolds et al. 1973). Another 20 years later, Dettmar (1990) and Rand et al. (1990) showed with deep Hα imaging of NGC891 that other galaxies also have such a warm, ionized component. A very complete and recent review of this so-called warm ionized medium (WIM) in galaxies has been given by Haffner et al. (2009).

The presence of dust in the ISM has been obvious for centuries: naked-eye observations of the central Milky Way show large dark patches of obscured starlight, bearing witness to the presence of obscuring material. Similar structures are also evident from images of other galaxies, where dust features outline the sometimes regular, sometimes chaotic spiral structure in galaxy disks. A major leap in studying dust in galaxies has come from infrared astronomy as noted in an early review (Stein and Soifer 1983) written at the dawn of routine mid-IR and far-IR imaging. A major breakthrough came from imaging of galaxies in the far-IR by IRAS (http://irsa.ipac.caltech.edu/IRASdocs/iras.html), and later the Infrared Space Observatory (ISO; http://iso.esac.esa.int), SPITZER (http://www.spitzer.caltech.edu), and HERSCHEL (http://www.esa.int/herschel). Not only detailed spectral energy distributions are now available, supporting the models describing the overall IR spectrum as thermal emission from cold dust, but also many spectral features resulting from complex molecules, primarily Polycyclic Aromatic Hydrocarbons (PAHs; see ➋ Chap. 10 of Volume 5) providing better insight in the interpretation of mid-IR and far-IR images of galaxies.

The tradition of observing H I and H II in galaxies is by now half a century old. The realization that molecules in the ISM can also be imaged using radio lines came almost two decades later with the first detection of CO in the Galaxy in 1970 (Wilson et al. 1970) and the first extragalactic CO detection in M 82 and NGC 253 in 1975 (Rickard et al. 1975a, b). The CO emission indirectly traces H_2, as rotational transitions of the CO molecules are excited by collisions with hydrogen molecules. Unfortunately, it is very difficult to observe the cold H_2 directly. Electronic transitions are very faint and occur in the ultraviolet where both the Earth's atmosphere and interstellar extinction obstruct our view (Shull and Beckwith 1982; Habart et al. 2005). Rotational transitions occur in the mid-infrared (Black and van Dishoeck 1987) at 28.2, 17.0, and 12.3 µm. H_2 can also be observed at shorter wavelengths when the temperature increases to several hundred K, but this only occurs around hot stars and active galactic nuclei (AGN). These lines of cold molecular hydrogen are weak and require very high column densities to be observable. So far, only detections of fairly warm gas have been made in the edge-on disk of NGC 891 (Valentijn and van der Werf 1999) and the nucleus of NGC 6946 (Valentijn et al. 1996).

Young and Scoville (1991) reviewed the early work on the CO and hence molecular content of galaxies, mostly carried out with single-dish radio telescopes. The first detailed images came with the advent of mm arrays and eventually large observing programs such as BIMA-SONG (Regan et al. 2001; Helfer et al. 2003) and HERACLES (Leroy et al. 2008, 2009).

A major uncertainty remains the conversion from CO brightness to H_2 column density. The general relation is expressed as

$$\frac{N_{H_2}}{\mathrm{cm}^{-2}} = X_{\mathrm{CO}} \int \frac{I_{\mathrm{CO}}(v)}{\mathrm{K\,kms}^{-1}}\, dv \tag{4.7}$$

The commonly adopted value for X_{CO} is the value determined for Galactic molecular clouds, $X_{\mathrm{CO}} = 3 \times 10^{20}\ \mathrm{cm}^{-2}\ (\mathrm{K\,km\,s}^{-1})^{-1}$. This conversion factor has, however, been demonstrated to depend on metallicity and can be uncertain by an order of magnitude (Wilson 1995; Israel 1997; Leroy et al. 2011a; see also Shetty et al. 2011a, b; Liszt et al. 2010 for a theoretical treatment of the X_{CO} conversion factor for both dense and diffuse molecular gas).

3 Star Formation and the ISM

The transformation of gas into stars is one of the most important processes in galaxy evolution. Understanding the conditions that determine the efficiency of this process, and the associated physics, is the goal of many observational and theoretical studies. They also form important input into numerical computer models of galaxy formation and evolution. A complete understanding requires knowledge of these processes over a large range in scales: from galaxy-sized scales where gas is transported from the disk of the galaxy into the halo and back (which is also the scale where accretion and capture of satellites takes place) to kpc-sized scales where gas clouds are coalescing, via sub-kpc scales where neutral gas cools and forms molecular gas in GMCs, to parsec scales where individual stars are formed.

These very small scales can be directly observed in our Galaxy, while the processes happening at galaxy scales can in principle be studied in external galaxies. Tying together the processes happening at these two extreme scales has been a major challenge in this particular field: in our Galaxy, we lack the overview we have for other galaxies, while we rarely have the resolution to study the detailed processes leading to star formation in external galaxies.

This led to many studies exploring empirical relations between the properties (usually surface or volume density) of the gas component and some kind of tracer of the recent star formation. In the last couple of years, significant progress has been made with the advent of high-resolution multiwavelength surveys of the gas and star formation components. These surveys are now starting to bridge the gap between the observations at the scales of stars and those at scales of galaxies.

Schmidt (1959) was the first to relate the gas density in our Galaxy to the star formation rate (SFR) using a relation $\rho_{\mathrm{SFR}} \sim \rho_{\mathrm{gas}}^{n}$. He found a value $n \simeq 2$. Schmidt's relation between volume densities can be translated into a relation involving surface densities $\Sigma_{\mathrm{SFR}} \sim \Sigma_{\mathrm{gas}}^{N}$. Surface densities are more easily quantifiable in external galaxies, and the exponents n and N are related as long as the scale height of the disk is constant. This kind of relation has become known as the "Schmidt Law," and the name is commonly used for any sort of relation that describes the link between a gas component (neutral, molecular, or global) and a star formation tracer (young stars, ionized gas emission, infrared emission).

The first investigations of the Schmidt Law in external galaxies were done by Sanduleak (1969) and Hartwick (1971), who looked at young stars and star-forming regions in the SMC and M31. Madore et al. (1974), Newton (1980), Tosa and Hamajima (1975), and Hamajima and Tosa (1975) followed up on this by studying the relation between young stars and HI surface density in nearby galaxies such as M31, the LMC, M101 and others. These studies found $N \simeq 2$ but with a large spread of a few tens of percent.

In a landmark paper, Kennicutt (1989) (see also Kennicutt 1998) studied the averaged gas and Hα content of 61 nearby spirals and 36 starburst galaxies. He derived a Schmidt Law between the total gas surface density $\Sigma_{\mathrm{gas}} = \Sigma_{HI} + \Sigma_{H_2}$ and the Hα-derived SFR with values of $N = 2.47 \pm 0.39$ for spirals and $N = 1.40 \pm 0.15$ for all galaxies (spirals and starbursts) in his sample. As a result of this work, the relation between gas surface density and star formation is often referred to as the "Kennicutt-Schmidt Law." ➜ *Figure 4-6* shows this result as presented in the Kennicutt (1989) paper.

Similar studies, using different measures for the gas surface density, and different star formation tracers (such as ultraviolet or infrared emission) found similar values, but again with a large spread. This may reflect actual variations in the physics, but a large part of this spread is very likely also due to choice of sample, analysis, and star formation tracers. For a description of these studies, see the review by Kennicutt (1998).

Kennicutt (1989, 1998) and Martin and Kennicutt (2001) also found evidence for a star formation threshold. That is, the star formation efficiency (as traced by Hα) decreases dramatically once the gas surface density drops below a certain threshold, but star formation does not shut off completely (e.g., Ferguson et al. 1998). Kennicutt (1989) relates this star formation threshold density to the Toomre Q parameter and the stability of the disk.

A similar cutoff behavior in the star formation efficiency had also been observed by Skillman (1987), who noted that star formation (again, as traced by Hα) did not seem to occur below a (constant) H I surface density threshold of $\sim 1 \cdot 10^{21}$ cm^{-2}. The star formation threshold was also suggested by van der Hulst et al. (1993) as an explanation for the low star formation rates observed in low surface brightness (LSB) galaxies. These galaxies generally have lower H I surface densities than "normal" galaxies (de Blok et al. 1996). Though this is not the place to review the large body of theoretical and numerical work on the relation between gas (surface) density and star formation rate, we note the work by Schaye (2004) and Taylor and Webster (2005) who suggest that the star formation threshold is related to the formation of a cold phase in the neutral ISM when a sufficiently high surface density is reached. This cold phase makes efficient cooling suddenly possible, leading to star formation.

Over the last decade, the amount of high-resolution, multiwavelength information about galaxies in the nearby universe has increased dramatically. This had made possible new studies of the conditions for star formation on kpc or even sub-kpc scales in a significant number of galaxies. Here we summarize some of the main results derived and refer to these papers for further references.

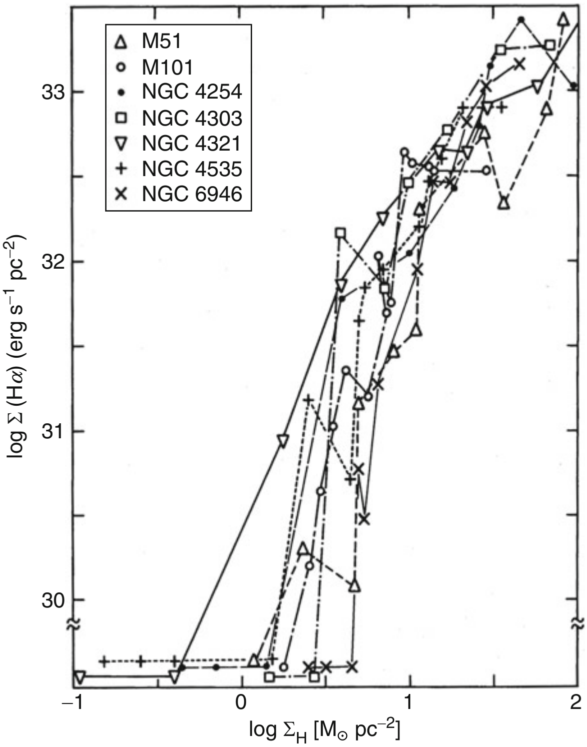


Fig. 4-6

Star formation surface density (measured from the Hα emission) as a function of gas surface density (measured from H I and H_2 (i.e., CO) observations) in seven well-resolved, nearby galaxies (Adopted from Kennicutt 1989)

Bigiel et al. (2008) investigated relations of the star formation rate surface density Σ_{SFR} with the HI surface density Σ_{HI}, the molecular surface density Σ_{H_2} and the combined gas surface density Σ_{gas} on scales of 750 pc in a number of nearby disk and dwarf galaxies. The star-formation rate maps were based on a combination of GALEX FUV maps and Spitzer 24 μm maps from respectively the GALEX NGS (Gil de Paz et al. 2007) and the Spitzer Infrared Nearby Galaxies Survey (SINGS; Kennicutt et al. 2003). For the gas surface densities, information was used from THINGS (Walter et al. 2008) for the neutral part of the ISM, and HERACLES (Leroy et al. 2009) and BIMA-SONG (Helfer et al. 2003) for the molecular part (as traced by the CO line). *Figure 4-7* shows three galaxies with such detailed information: the distribution of molecular hydrogen, atomic hydrogen, the kinematics of the atomic hydrogen, the distribution of near-IR light (predominantly old stars), and a composite of UV light, mid-IR light, and Hα emission (tracing star formation).

It was found that, at least in the star-forming disks of the sample galaxies, a strong relation exists between Σ_{H_2} and Σ_{SFR}, with a power-law exponent $N = 1.0 \pm 0.2$, i.e., a linear relation. This can be interpreted as stars forming from the molecular ISM at a constant efficiency. This agrees with a result from Kennicutt (1989) who had found a similar relation for starburst galaxies,

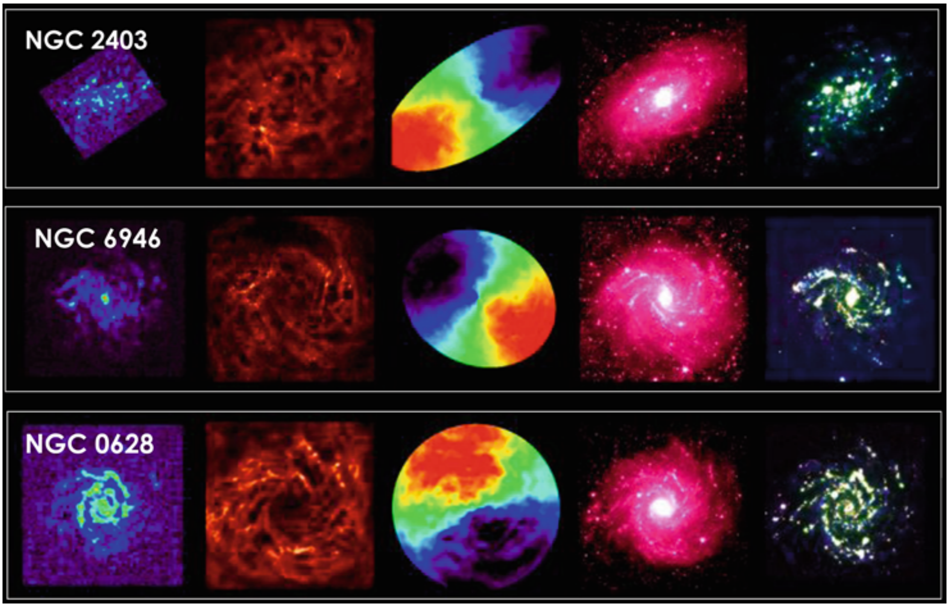


Fig. 4-7
Three examples of the wealth of information available to study the relation between gas density, gas kinematics, and star formation in detail (courtesy Adam Leroy and the HERACLES team, Leroy et al. 2012, 2011b). From *left to right*: distribution of CO (as proxy for H_2) (HERACLES), distribution of H I (THINGS), H I velocity field, near-IR image (SPITZER: SINGS and LVL), and composite image from GALEX FUV, SINGS/LVL mid-IR, and SINGS/LVL Hα emission showing the distribution of star formation

where the H_2 surface densities are much higher. This is illustrated in *Fig. 4-8* (adopted from Fig. 10 of Schruba et al. 2011), which shows two different measures of the star formation surface density versus the gas surface density. The power-law regime applies to gas surface densities above 10 $M_\odot$ pc^{-2}.

Note, though, that the Bigiel et al. (2008) result does not say anything about the necessary conditions for star formation inside GMCs. At the working resolution of 750 pc, GMCs are not resolved, and the linear relation found must therefore be interpreted within the context of "counting clouds," i.e., they can be explained by assuming that the GMCs have uniform properties with the observed CO surface density determined by the beam filling factor of these clouds.

Bigiel et al. (2008) found less well-defined relations between $\Sigma_{\rm gas}$ or Σ_{HI} on the one hand and $\Sigma_{\rm SFR}$ on the other hand. The relation between $\Sigma_{\rm gas}$ and $\Sigma_{\rm SFR}$ was found to vary from galaxy to galaxy, and there is almost no relation with Σ_{HI}. These results were revisited by Schruba et al. (2011) who used stacking of the HERACLES CO data (using HI radial velocities as a prior) to push the CO detections to larger radii. In these outer parts, the gas content is dominated by the neutral HI rather than the molecular H_2. In this lower gas surface density regime, the star formation law changes its linear behavior. This is clear in *Fig. 4-8*.

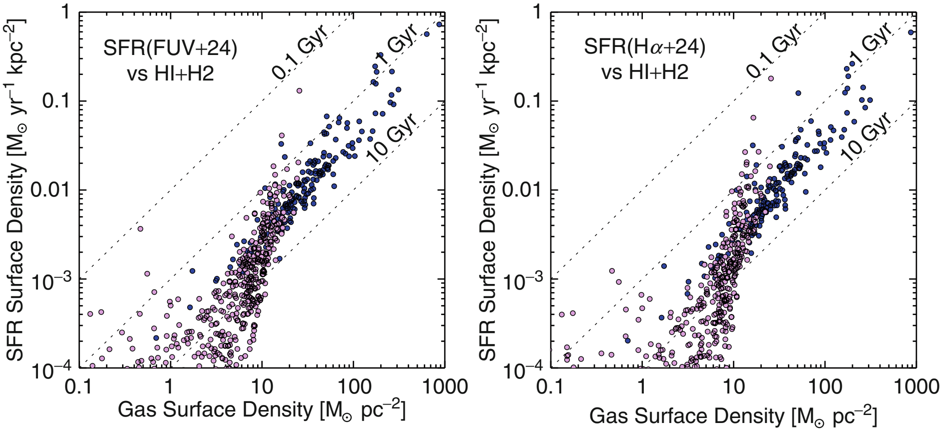


Fig. 4-8
Σ_{SFR} (y-axis) from FUV + 24 μm (*left*) and Hα + 24 μm (*right*) as a function of gas surface density (H I + H₂). Each *point* represents a radial average in a given galaxy. Regions that are H_2 dominated are plotted with *dark blue symbols*, regions that are H I dominated in *light red symbols*. Whereas SFR is not correlated with H I (in the inner parts of galaxy disks), it correlates with H_2 and total gas surface density. The scaling exhibits a change in slope at the transition between H I- and H_2-dominated environments (This figure is adopted from Fig. 11 in Schruba et al. 2011)

Schruba et al. (2011) found that the linear relation between Σ_{H_2} and $\Sigma_{\rm SFR}$ extends into this HI-dominated regime. One interpretation of these results is that the Kennicutt-Schmidt Law can be separated into two separate processes. The first one is star formation proceeding with a constant efficiency once H_2 is formed. The "bottleneck" determining the star formation rate thus seems to be the second process of conversion from HI into H_2 (likely through the formation of a cold neutral component as discussed earlier). The star-formation threshold as found by Kennicutt (1989) therefore seems to be caused by a change in SFR due to a changing ratio of HI to H_2 (as a function of total gas density) rather than an actual shutting off of star formation.

This is supported by the realization that star formation in the outer parts of disks is more widespread than originally thought. The early observations of Ferguson et al. (1998) showed isolated star-forming regions (as traced by Hα) well beyond the optical disks of a number of nearby galaxies. The discovery of further star formation using GALEX observations in the UV showed that these early observations were showing only the tip of the iceberg (Thilker et al. 2005). The UV observations enabled direct detection of O and B stars which would otherwise have escaped detection due to their inability to excite the surrounding ISM enough to produce Hα emission. It is now thought that these so-called extended UV (XUV) disks are found in ~30% of nearby disk galaxies (Thilker et al. 2007). A striking example is M 83 (NGC 5236) shown in ➲ *Fig. 4-9*: the outlying H I structures show up remarkably well in the UV, indicating that star formation is progressing there as well, albeit with a lower efficiency (Bigiel et al. 2010b) than in the bright, inner region of the galaxy. A comparison of star formation as traced by Hα with star formation as traced by UV radiation from O and B stars shows that while the level of Hα emission typically drops very steeply at the edge of the optical disk (giving the appearance of a star formation threshold), such an edge is not observed in the UV emission. Rather, the

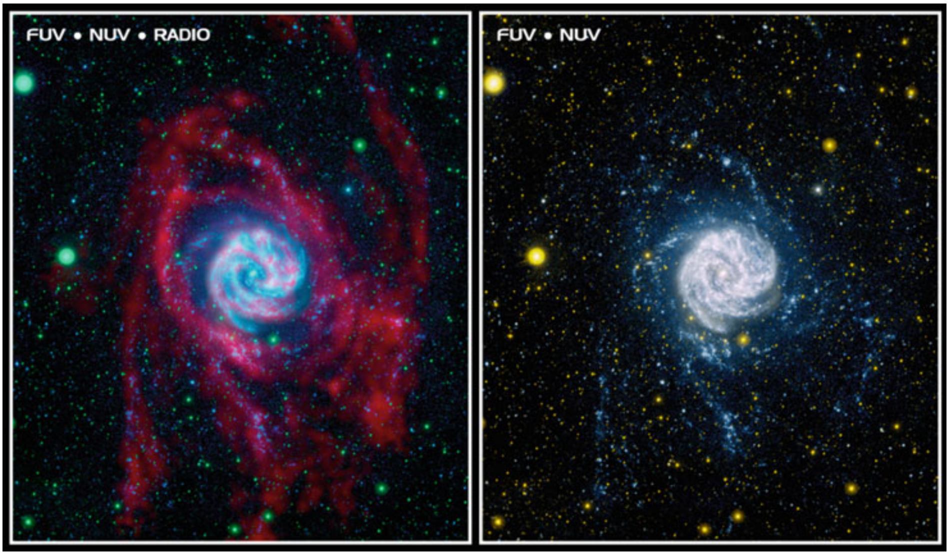


Fig. 4-9
THINGS H I image (Walter et al. 2008) superposed on a near- and far-UV composite image (Thilker et al. 2005; Bigiel et al. 2010a) of M 83 (*left panel*). For better comparison of the faint outer star formation structures the single UV composite image is shown as well (*right panel*). Note the extended structures, (arms, arcs, filaments) in the outer disk, visible in both the H I and the UV images, indicating that even in the faint outer H I star formation is progressing (Courtesy Dave Thilker and NASA/JPL-Caltech/VLA/MPIA)

density of young stars gradually decreases, again supporting the idea that the changing ratio of HI to H_2 ultimately determines the overall star formation efficiency.

The above broad-brush picture illustrates the progress made in the last few years, and these results can now be used as input for numerical models that can help explain variations of star formation efficiency with, e.g., redshift, environment, or galaxy mass.

It is, however, still difficult to link the observations with the actual physical processes driving the star formation rate. Leroy et al. (2008) used the same data set as Bigiel et al. (2008) to test a number of theoretical explanations proposed in the literature linking the gas density and the SFR. They looked at the disk free-fall time, the orbital timescale, the effects of cloud-cloud collisions, the assumption of fixed GMC star formation efficiency, and the relation between pressure in the ISM and the phases of the ISM.

Similarly, they investigated several models proposed for the star formation threshold, such as gravitational stability in a gas disk, gravitational stability in a mixed disk of gas and stars, the effects of shear in a disk, and the onset of a cold gas phase.

Their conclusions were that none of these offers a unique explanation for the observed behavior. While large-scale relations can be identified (such as the Kennicutt-Schmidt Law), the actual physics remains more complicated, and still below the resolutions achieved so far. Observationally we will also have to get a better understanding of the balance between warm and cold HI phases, the efficiency of H_2 formation, and possibly the effects of shocks and turbulence. Many of these have already been studied numerically or theoretically. In the next decade or so we, should be able to obtain the observational evidence in a large number of galaxies at sufficient resolution in order to gauge the ability of the ISM to form GMCs over a wide range of galaxy conditions.

4 Accretion, Feedback and the Environment

One of the main puzzles in galaxy evolution is that the rate at which new stars have formed in galaxies has declined dramatically over the last 7 Gyr (e.g., Madau et al. 1998; Hopkins and Beacom 2006), in contrast to the lack of a similar decline in the estimated cold gas density in the universe (Lah et al. 2007). Observed star formation rates in galaxies are such that the observed gas supply is exhausted in a few Gyr (e.g., Bigiel et al. 2008) and the dwindling star formation could, in principle, be due to star formation consuming the available gas supply. Galaxies must, therefore, continuously accrete gas from the intergalactic environment to keep the observed gas density levels. Continuous accretion of gas from the IGM may solve this discrepancy so that star formation can continue over much longer periods than given by the initial gas reservoirs and the gas consumption times.

The observational evidence for accretion has been extensively described in the review of Sancisi et al. (2008), who examine a number of H I signatures in galaxies in the nearby universe. Examples are small (and often subtle) asymmetries in the distribution and/or kinematics of the H I (lopsidedness), the presence of warps, of small H I companions, faint H I tails and other structures, and the presence of extraplanar gas (a prime example being NGC 891, Oosterloo et al. (2007a)). ➲ *Figure 4-10* provides a few examples of galaxies with this kind of evidence for accretion. From an inventory of such H I signatures in WHISP (van der Hulst et al. 2001; Sancisi et al. 2008) estimate an accretion rate of ~0.2 $M_\odot$year^{-1} in H I. This number is very uncertain and does not account for the accretion of warm, ionized gas.

From the theoretical point of view, there is increasing evidence that so-called cold-mode accretion plays an important role in providing the dark matter halos with the fuel for forming stars and building up the stellar mass in galaxies (Birnboim and Dekel 2003; Dekel and Birnboim 2006; Binney 2004; Kereš et al. 2005; Cattaneo et al. 2006; Kereš et al. 2009a). Cosmological simulations suggest that cold-mode accretion is the dominant process at redshifts $z \geq 2$ but becomes less important at lower redshifts (van de Voort et al. 2011). The cold-mode accretion has been examined closely in the Milky Way and a few nearby galaxies (NGC 891 and NGC 2403) by Fraternali and Binney (2008), Marinacci et al. (2010) and Binney and Fraternali (2012), in particular with the goal of verifying theoretically what fraction of the gas in the halo of the Milky Way results from accretion versus gas brought into the halo by the "Galactic Fountain" mechanism. Marasco and Fraternali (2011) and Marasco et al. (2012) conclude from detailed modeling of the halo of the Milky Way that the "Galactic Fountain" induces a steady inflow of ~2 $M_\odot$ year^{-1}, enough to sustain the star formation in the disk and an order of magnitude larger than the accretion rate estimated by Sancisi et al. (2008).

Accretion is not the only process relevant to the interplay between star formation and the environment. Gas removal processes can be equally important depending on internal processes (stellar winds, supernovae, presence of an AGN) or external, environmental conditions (density of the intergalactic medium (IGM), local galaxy density, i.e., the probability of gravitational interactions). Examples of such external processes are plentiful in the denser environments of clusters and groups. The best observational evidence is found in the Virgo cluster (➲ *Fig. 4-11*), which offered the first clear effects of gas removal by ram pressure in the hot intracluster gas and gravitational interactions (Warmels 1988; Cayatte et al. 1994; Chung et al. 2009). These studies

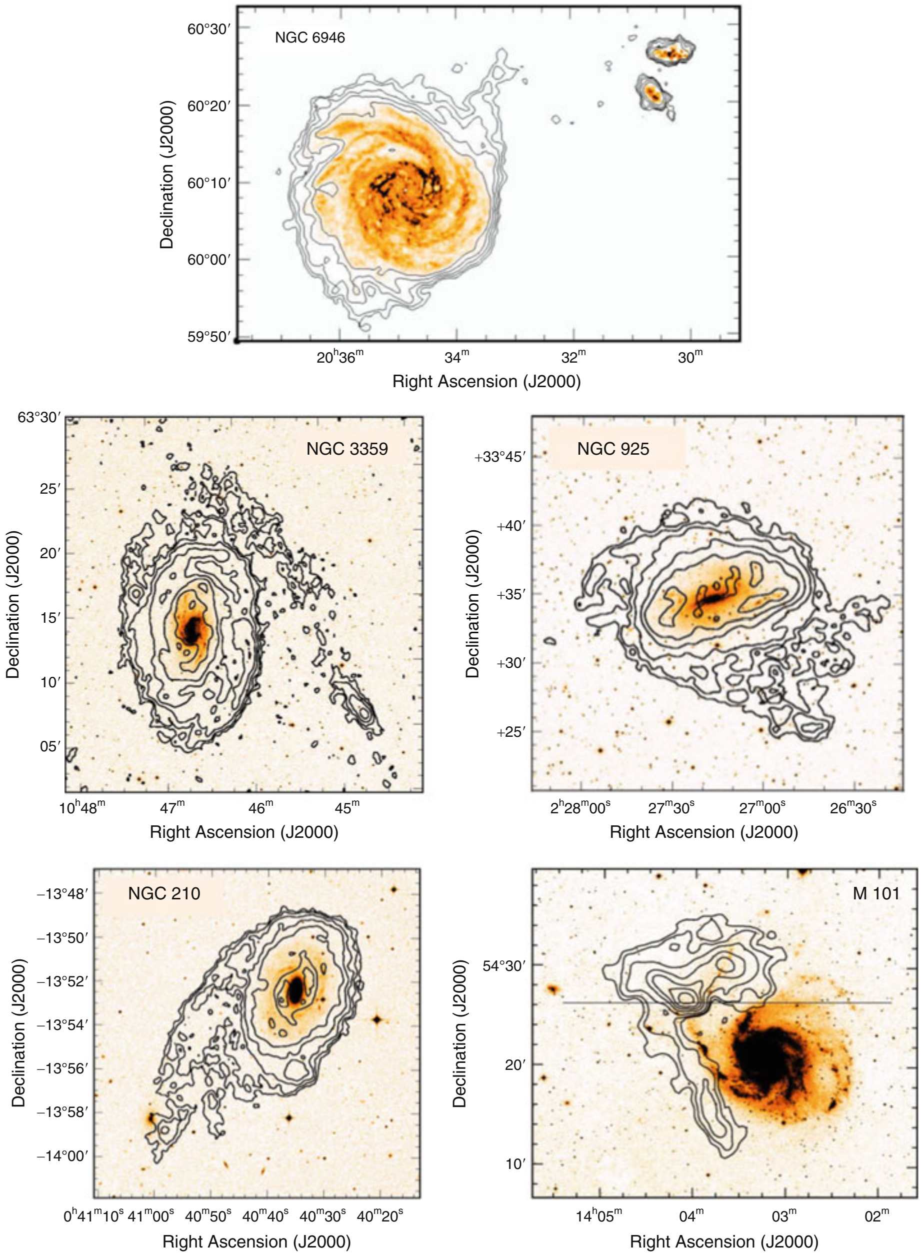


Fig. 4-10

Five examples of H I structures which could be the result of recent interactions/accretion. In each panel, the contours show the H I density distribution superposed on the optical image. For NGC 6946, the column density levels are 1.25, 2.5, 5, 10, and 20 × 10^{19} cm^{-2}; for NGC 3359, 10, 20, 50, 100, 200, 400 × 10^{19} cm^{-2}; for NGC 925, 5, 10, 20, 50, 100 × 10^{19} cm^{-2}; for NGC 210, 5, 10, 20, 50, 100, 200 × 10^{19} cm^{-2}; and for M 101, 0.7, 1.4, 2.8, 5.6, 8.4, 11.2, 14.0 × 10^{19} cm^{-2}. In M101 only the high velocity gas is shown. (Images are courtesy of Oosterloo and Sancisi, and from Boomsma et al. 2008 and Kamphuis 1993)

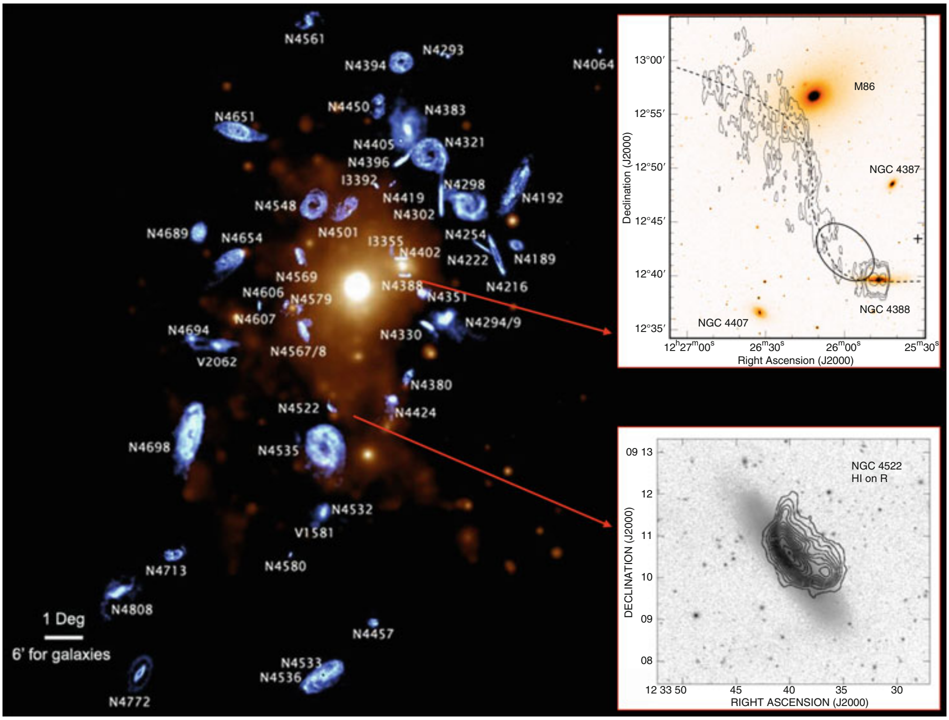


Fig. 4-11

Examples of effects of the environment in the Virgo cluster. The *right panel* shows the central part of the Virgo cluster with in red the X-ray emission from the intracluster gas and individual massive galaxies and in blue the H I distributions of the galaxies imaged with the VLA by Chung et al. (2009). Note that the H I disks are enlarged by a factor 10 for visible presentation. The H I disks in the core of the Virgo cluster are clearly smaller than those of galaxies in the outer parts. To the *right*, two particular examples of ram pressure stripping are enlarged: NGC 4388 (Oosterloo and van Gorkom 2005) and NGC 4522 (Kenney et al. 2004)

showed that all galaxies are systematically small in H I in the cluster core where the IGM density is highest (➋ *Fig. 4-11*), with two clear examples of ongoing ram pressure stripping: NGC 4522 (Kenney et al. 2004), and NGC 4388 (Oosterloo and van Gorkom 2005). These studies provided insight into why the relative H I content in galaxies in dense environments is low (Solanes et al. 2001).

Internal gas removal processes, often referred to as "feedback," are most obvious in galaxies with either an AGN or a massive starburst. Very telling examples are the nearby galaxies M 82 and NGC 253. ➋ *Figure 4-12* shows the H I, the ionized gas and the hot X-ray gas blown out of the central disk by the massive nuclear starburst in NGC 253 (Boomsma et al. 2005). For a review of such massive starbursts, see Heckman et al. (1993). These phenomena are much more energetic than the moderate "Galactic Fountain" mechanism but may have much shorter duration. The combination of such processes has been implemented in theoretical models describing galaxy evolution (Cook et al. 2010; Kereš et al. 2009a, b; Springel et al. 2005) and

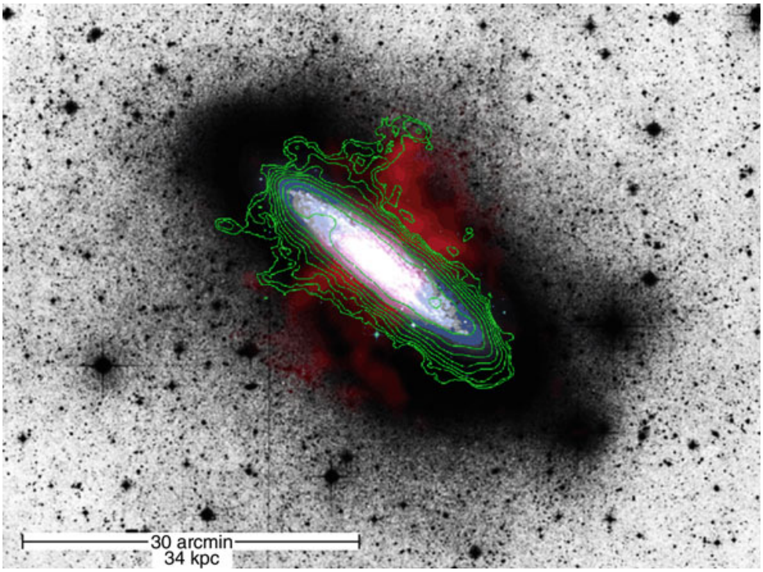


Fig. 4-12
Deep optical image of NGC 253 with superposed H I column density contours (in green at 0.18, 0.36, 0.72, 1.4, 2.9, 5.8, 12, and 23 × 10^{20} cm^{-2}) and 0.1–0.4 keV X-ray emission (in *red*). The nuclear outburst is clearly visible in X-rays outlining the hot gas, and in H I outlining the cooler gas at the periphery of the nuclear wind (Boomsma et al. 2005)

appears to be important for regulating the star formation such that the observed luminosity and H I mass functions can be reproduced.

Understanding the balance between gas removal and gas accretion processes, including the role of the environment, will be a crucial element of understanding galaxy evolution. Though the current work on H I in a variety of environments is beginning to provide insight (van Gorkom 2008, 2011), it is based on only a small number of cases. New facilities, recently or soon available, will image an order of magnitude more galaxies and therefore contribute significantly to determining this balance. Such facilities are ALMA, the E-VLA, APERTIF on the WSRT (Verheijen et al. 2008, 2009; Oosterloo et al. 2010b), MeerKAT (de Blok et al. 2010; de Blok 2011), and ASKAP (Johnston et al. 2009; Westmeier and Johnston 2010).

References

Baldwin, J. E., Field, C., Warner, P. J., & Wright, M. C. H. 1971, MNRAS, 154, 445

Battaglia, G., Fraternali, F., Oosterloo, T., & Sancisi, R. 2005, A&A, 447, 49

Begum, A., Chengalur, J. N., & Karachentsev, I. D. 2005, A&A, 433, L1

Begum, A., Chengalur, J. N., Karachentsev, I. D., Sharina, M. E., & Kaisin, S. S. 2008, MNRAS, 386, 1667 (FIGGS)

Bigiel, F., Leroy, A., Walter, F., Brinks, E., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2846 (THINGS)

Bigiel, F., Leroy, A., Seibert, M., Walter, F., Blitz, L., Thilker, D., & Madore, B. 2010, ApJ, 720, L31

Bigiel, F., Leroy, A., Walter, F., Blitz, L., Brinks, E., de Blok, W. J. G., & Madore, B. 2010, AJ, 140, 1194

Binney, J. 2004, MNRAS, 347, 1093

Binney, J., & Fraternali, F. 2012, EPJWC, 19, 8001

Birnboim, Y., & Dekel, A. 2003, MNRAS, 345, 349
Black, J. H., & van Dishoeck, E. F. 1987, ApJ, 322, 412
Blitz, L., & Rosolowsky, E. 2006, ApJ, 650, 933
Boomsma, R., Oosterloo, T. A., Fraternali, F., van der Hulst, J. M., & Sancisi, R. 2005, A&A, 431, 65
Boomsma, R., Oosterloo, T. A., Fraternali, F., van der Hulst, J. M., & Sancisi, R. 2008, A&A, 490, 555
Bregman, J. N. 1980, ApJ, 236, 577
Briggs, F. H. 1990, ApJ, 352, 15
Broeils, A. H., & Rhee, M.-H. 1997, A&A, 324, 877
Cattaneo, A., Dekel, A., Devriendt, J., Guiderdoni, B., & Blaizot, J. 2006, MNRAS, 370, 1651
Cayatte, V., Kotanyi, C., Balkowski, C., & van Gorkom, J. H. 1994, AJ, 107, 1003
Chung, A., van Gorkom, J. H., Kenney, J. D. P., Crowl, H., & Vollmer, B. 2009, AJ, 138, 1741
Cook, M., Barausse, E., Evoli, C., Lapi, A., & Granato, G. L. 2010, MNRAS, 402, 2113
de Blok, W. J. G., McGaugh, S. S., & van der Hulst, J. M. 1996, MNRAS, 283, 18
de Blok, W. J. G. 2011, IAU Symp., 277, 96
de Blok, E. W. J. G., Booth, R., Jonas, J., & Fanaroff, B. 2010, PoS (ISKAF2010) 005, http://pos.sissa.it/archive/conferences/112/005/ISKAF2010_005.pdf
Dekel, A., & Birnboim, Y. 2006, MNRAS, 368, 2
Dettmar, R.-J. 1990, A&A, 232, L15
Ewen, H. I., & Purcell, E. M. 1951, Nature, 168, 356
Ferguson, A. M. N., Wyse, R. F. G., Gallagher, J. S., & Hunter, D. A. 1998, ApJ, 506, 19L
Field, G. B. 1959, ApJ, 129, 536
Fraternali, F., Oosterloo, T., Sancisi, R., & van Moorsel, G. 2001, ApJ, 562, 47
Fraternali, F., van Moorsel, G., Sancisi, R., & Oosterloo, T. 2002, AJ, 123, 3124
Fraternali, F., & Binney, J. J. 2008, MNRAS, 386, 935
García-Ruiz, I., Sancisi, R., & Kuijken, K. 2002, A&A, 394, 769 (WHISP)
Gil de Paz, A., et al. 2007, ApJS, 173, 185
Giovanelli, R., et al. 2005, AJ, 130, 2598 and 2613 (ALFALFA)
Habart, E., Walmsley, M., Verstraete, L., Cazaux, S., Maiolino, R., Cox, P., Boulanger, F., & Pineau des Forêts, G. 2005, Space Sci. Rev., 119, 71
Haffner, L. M., Dettmar, R.-J., Beckman, J. E., et al. 2009, Rev. Mod. Phys., 81, 969
Hamajima, K., & Tosa, M. 1975, PASJ, 27, 561
Hartwick, F. D. A. 1971, ApJ, 163, 431
Heald, G., et al. 2011, A&A, 526, A118 (HALOGAS)
Heckman, T. M., Lehnert, M. D., & Armus, L. 1993, The environment and evolution of galaxies. Astrophys. Space Sci. Libr., 188, 455
Helfer, T. T., Thornley, M. D., Regan, M. W., Wong, T., Sheth, K., Vogel, S. N., Blitz, L., & Bock, D. C.-J. 2003, ApJS, 145, 259 (BIMA-SONG)
Hopkins, A. M., & Beacom, J. F. 2006, ApJ, 651, 142
Hoyle, F., & Ellis, G. R. A. 1963, Aust. J. Phys., 16, 1
Hunter, D. A., Elmegreen, B. G., Oh, S.-H., et al. 2011, AJ, 142, 121 (LITTLE-THINGS)
Israel, F. P. 1997, A&A, 328, 471
Johnston, S., Feain, I. J., & Gupta, N. 2009, ASPC, 407, 446
Józsa, G. I. G., Kenn, F., Klein, U., & Oosterloo, T. A. 2007, A&A, 468, 731
Józsa, G. I. G. 2007, A&A, 468, 903
Kahn, F. D. 1994, Ap&SS, 216, 325
Kalberla, P. M. W., & Kerp, J. 2009, ARA&A, 47, 27
Kamphuis, J. J. 1993, PhD Thesis, University of Groningen
Kamphuis, J., & Sancisi, R. 1993, A&A, 273, 31
Kenney, J. D. P., van Gorkom, J. H., & Vollmer, B. 2004, AJ, 127, 3361
Keres, D., Katz, N., Weinberg, D. H., & Davé, R. 2005, MNRAS, 363, 2
Kereš, D., Katz, N., Fardal, M., Davé, R., & Weinberg, D. H. 2009a, MNRAS, 395, 160
Kereš, D., Katz, N., Davé, R., Fardal, M., & Weinberg, D. H. 2009b, MNRAS, 396, 2332
Kerr, F. J. 1969, ARA&A, 7, 39
Kennicutt, R. C., Jr. 1989, ApJ, 344, 685
Kennicutt, R. C., Jr. 1998, ARA&A, 36, 189
Kennicutt, R. C., Jr., et al. 2003, PASP, 115, 928
Koribalski, B. S., et al. 2004, AJ, 128, 16 (HIPASS)
Kregel, M., & van der Kruit, P. C. 2005, MNRAS, 358, 481
Lah, P., et al. 2007, MNRAS, 376, 1357
Leroy, A. K., Walter, F., Brinks, E., Bigiel, F., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2782 (HERACLES) (THINGS)
Leroy, A. K., et al. 2009, AJ, 137, 4670 (HERACLES) (THINGS)
Leroy, A. K., et al. 2011a, ApJ, 737, 12 (HERACLES) (THINGS)
Leroy, A. K., Walter, F., Schruba, A., Bigiel, F., Foyle, K., & HERACLES Team 2011b, AAS, 43, #246.14
Leroy, A. K., Walter, F., Schruba, A., & HERACLES Collaboration 2012, AAS, 219, #346.03
Levine, E. S., Blitz, L., & Heiles, C. 2006, ApJ, 643, 881
Liszt, H. S., Pety, J., & Lucas, R. 2010, A&A, 518, A45
Madau, P., Pozzetti, L., & Dickinson, M. 1998, ApJ, 498, 106
Madore, B. F., van den Bergh, S., & Rogstad, D. H. 1974, ApJ, 191, 317
Marasco, A., & Fraternali, F. 2011, A&A, 525, A134
Marasco, A., Fraternali, F., & Binney, J. J. 2012, MNRAS, 419, 1107
Marinacci, F., Binney, J., Fraternali, F., Nipoti, C., Ciotti, L., & Londrillo, P. 2010, MNRAS, 404, 1464
Martin, C. L., & Kennicutt, R. C., Jr. 2001, ApJ, 555, 301
Martínez-Delgado, D., et al. 2010, AJ, 140, 962

Meyer, M. J., et al. 2004, MNRAS, 350, 1195 (HIPASS)

Muller, C. A., & Oort, J. H. 1951, Nature, 168, 357

Newton, K. 1980, MNRAS, 190, 689

Noordermeer, E., van der Hulst, J. M., Sancisi, R., Swaters, R. A., & van Albada, T. S. 2005, A&A, 442, 137 [WHISP]

Norman, C. A., & Ikeuchi, S. 1989, ApJ, 345, 372

O'Brien, J. C., Freeman, K. C., & van der Kruit, P. C. 2010, A&A, 515, A62 and A63

Olling, R. P. 1996, AJ, 112, 457

Oosterloo, T., & van Gorkom, J. 2005, A&A, 437, L19

Oosterloo, T., Fraternali, F., & Sancisi, R. 2007a, AJ, 134, 1019

Oosterloo, T. A., et al. 2007b, NewAR, 51, 8

Oosterloo, T. A., Morganti, R., Sadler, E. M., van der Hulst, T., & Serra, P. 2007c, A&A, 465, 787

Oosterloo, T. A., et al. 2010a, MNRAS, 409, 500

Oosterloo, T., Verheijen, M., & van Cappellen, W. 2010b, PoS (ISKAF2010) 043 http://pos.sissa.it/archive/conferences/112/043/ISKAF2010_043.pdf

Ott, J., Skillman, E., Dalcanton, J., Walter, F., Stilp, A., Koribalski, B., West, A., & Warren, S. 2008, AIPC, 1035, 105 (VLA-ANGST)

Pawsey, J. L. 1951, Nature, 168, 358

Putman, M. E., Staveley-Smith, L., Freeman, K. C., Gibson, B. K., & Barnes, D. G. 2003, ApJ, 586, 170

Raimond, E., & Volders, L. M. J. S. 1957, BAN, 14, 19

Rand, R. J., Kulkarni, S. R., & Hester, J. J. 1990, ApJ, 352, L1

Regan, M. W., Thornley, M. D., Helfer, T. T., Sheth, K., Wong, T., Vogel, S. N., Blitz, L., & Bock, D. C.-J. 2001, ApJ, 561, 218 (BIMA-SONG)

Reynolds, R. J. 1971, PhD Thesis, University of Wisconsin

Reynolds, R. J., Scherb, F., & Roesler, F. L. 1973, ApJ, 185, 869

Rickard, L. J., Palmer, P., Morris, M., Zuckerman, B., & Turner, B. E. 1975a, BAAS, 7, 253

Rickard, L. J., Palmer, P., Morris, M., Zuckerman, B., & Turner, B. E. 1975b, BAAS, 7, 529

Roberts, M. S. 1963, ARA&A, 1, 149

Roberts, M. S. 1975, Stars and Stellar Systems, Vol. IX (Chicago: University of Chicago Press), 309

Roberts, M. S., & Haynes, M. P. 1994, ARA&A, 32, 115

Rogstad, D. H., & Shostak, G. S. 1971, A&A, 13, 99

Rosen, A., & Bregman, J. N. 1995, ApJ, 440, 634

Roychowdhury, S., Chengalur, J. N., Begum, A., & Karachentsev, I. D. 2009, MNRAS, 397, 1435 (FIGGS)

Roychowdhury, S., Chengalur, J. N., Begum, A., & Karachentsev, I. D. 2010, MNRAS, 404, L60 (FIGGS)

Rupen, M. P. 1991, AJ, 102, 48

Sancisi, R., & Allen, R. J. 1979, A&A, 74, 73

Sancisi, R., Fraternali, F., Oosterloo, T., & van der Hulst, T. 2008, A&AR, 15, 189

Sanders, R. H. 2010, The Dark Matter Problem (Cambridge University Press), ISBN:9780521113014

Sanduleak, N. 1969, AJ, 74, 47

Savage, B. D., & de Boer, K. S. 1979, ApJ, 230, L77

Schaap, W. E., Sancisi, R., & Swaters, R. A. 2000, A&A, 356, 49L

Schaye, J. 2004, ApJ, 609, 667

Schruba, A., et al. 2011, AJ, 142, 37

Schmidt, M. 1959, ApJ, 129, 243

Serra, P., et al. 2012, MNRAS, 422, 1835 (ATLAS3D)

Shetty, R., Glover, S. C., Dullemond, C. P., Ostriker, E. C., Harris, A. I., & Klessen, R. S. 2011, MNRAS, 415, 3253

Shetty, R., Glover, S. C., Dullemond, C. P., & Klessen, R. S. 2011, MNRAS, 412, 1686

Shull, J. M., & Beckwith, S. 1982, ARA&A, 20, 163

Skillman, E. D. 1987, NASCP, 2466, 263

Solanes, J. M., Manrique, A., García-Gómez, C., González-Casado, G., Giovanelli, R., & Haynes, M. P. 2001, ApJ, 548, 97

Springel, V., Di Matteo, T., & Hernquist, L. 2005, MNRAS, 361, 776

Stein, W. A., & Soifer, B. T. 1983, ARA&A, 21, 177

Swaters, R. A., Sancisi, R., & van der Hulst, J. M. 1997, ApJ, 491, 140

Swaters, R. A., van Albada, T. S., van der Hulst, J. M., & Sancisi, R. 2002, A&A, 390, 829 (WHISP)

Taylor, E. N., & Webster, R. L. 2005, ApJ, 634, 1067

Thilker, D. A., et al. 2005, ApJ, 619, L79

Thilker, D. A., et al. 2007, ApJS, 173, 538

Toribio, M. C., Solanes, J. M., Giovanelli, R., Haynes, M. P., & Martin, A. M. 2011, ApJ, 732, 93 [ALFALFA]

Tosa, M., & Hamajima, K. 1975, PASJ, 27, 501

Valentijn, E. A., & van der Werf, P. P. 1999, ApJ, 522, L29

Valentijn, E. A., van der Werf, P. P., de Graauw, T., & de Jong, T. 1996, A&A, 315, L145

van de Hulst, H. C. 1945, Ned. Tijdschr. voor Natuurkunde, 11, 201

van der Hulst, J. M., & Sancisi, R. 1988, AJ, 95, 1354

van der Hulst, J. M., Skillman, E. D., Smith, T. R., Bothun, G. D., McGaugh, S. S., & de Blok, W. J. G. 1993, AJ, 106, 548

van der Hulst, J. M., van Albada, T. S., & Sancisi, R. 2001, ASPC, 240, 451 (WHISP)

van der Kruit, P. C., & Freeman, K. C. 2011, ARA&A, 49, 301

van de Voort, F., Schaye, J., Booth, C. M., & Dalla Vecchia, C. 2011, MNRAS, 415, 2782

van Eymeren, J., Jütte, E., Jog, C. J., Stein, Y., & Dettmar, R.-J. 2011a, A&A, 530, A29 van Eymeren, J., Jütte, E., Jog, C. J., Stein, Y., & Dettmar, R.-J. 2011b, A&A, 530, A30

van Gorkom, J. 2008, AIPC, 1035, 24

van Gorkom, J. 2011, IAU Symp., 277, 41

Verheijen, M. A. W., Oosterloo, T. A., van Cappellen, W. A., Bakker, L., Ivashina, M. V., & van der Hulst, J. M. 2008, AIPC, 1035, 265

Verheijen, M., Oosterloo, T., Heald, G., & van Cappellen, W. 2009, PoS (PRA2009) 089, http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=89

Volders, L. M. J. S. 1959, BAN, 14, 323

Volders, L. M. J. S., & Högbom, J. A. 1961, BAN, 15, 307

Walter, F., Brinks, E., de Blok, W. J. G., Bigiel, F., Kennicutt, R. C., Jr., Thornley, M. D., & Leroy, A. 2008, AJ, 136, 2563 (THINGS)

Walterbos, R. A. M., & Braun, R. 1996, The Minnesota lectures on extragalactic neutral hydrogen. ASPC, 106, 1

Warmels, R. H. 1988, A&AS, 72, 19, 17, and 427

Westmeier, T., & Johnston, S. 2010, PoS (ISKAF2010) 056, http://pos.sissa.it/archive/conferences/112/056/ISKAF2010_056.pdf

Wilson, C. D. 1995, ApJ, 447, 616

Wilson, R. W., Jefferts, K. B., & Penzias, A. A. 1970, ApJ, 161, L43

Wright, M. C. H., Warner, P. J., & Baldwin, J. E. 1972, MNRAS, 155, 337

Young, J. S., & Scoville, N. Z. 1991, ARA&A, 29, 581

Zwaan, M. A., et al. 2004, MNRAS, 350, 1210 (HIPASS)

5 The Influence of Environment on Galaxy Evolution

Bernd Vollmer
Observatoire astronomique de Strasbourg, Strasbourg, France

1 *Introduction* 209

2 *Galaxy Populations in Groups and Clusters* 210

3 *Interaction Types* 211
3.1 Gravitational Interactions 211
3.2 Hydrodynamical Interactions 213

4 *Simulations* 215
4.1 Spherical Galaxies 216
4.2 Disk Galaxies 216

5 *The Multiphase ISM* 219
5.1 Atomic Hydrogen 219
5.2 Molecular Hydrogen 221
5.3 Dust and Metallicity 223
5.4 Cosmic Ray Gas 224

6 *Star Formation* 225

7 *The Global Picture* 226

8 *Resolved Multiwavelength Interaction Diagnostics* 231
8.1 Environmental Effects in Nearby Galaxy Groups 232
8.2 Environmental Effects in Nearby Galaxy Clusters 235
8.2.1 Early Type Galaxies 237
8.2.2 Late Type Galaxies 239
8.2.3 Dwarf Galaxies 245
8.3 A Holistic View on Ram Pressure Stripping 247
8.3.1 The Response of the Multiphase ISM and Star Formation to Ram Pressure 251

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_5, © Springer Science+Business Media Dordrecht 2013

9 The Detailed Picture 253

10 A Local View on the Butcher–Oemler Effect 255

11 Conclusions and Outlook 257

Acknowledgments 258

References 258

Abstract: Galaxy evolution is influenced by environment. The properties in terms of morphology, color, gas content, and star formation of galaxies residing in the field, groups, or clusters are markedly different. Environmental effects include gravitational interactions with other galaxies or the cluster potential and hydrodynamical effects as ram pressure stripping. An overview of the theoretical and observational aspects of galaxy evolution in different environments is given. Spherical, disk, and dwarf galaxies are discussed separately. Different simulation techniques for the modeling of environmental effects on the ISM are presented and compared. Environmental interactions leave imprints on the atomic and molecular hydrogen, dust, cosmic ray gas, and large-scale magnetic fields. They also modify the star formation of a galaxy that enters an environment of higher density. A global picture of galaxy evolution in different environments is drawn by combining integrated and resolved observations at multiple wavelengths. Special attention is given to multiwavelength interaction diagnostics of individual cluster galaxies. This leads to a more detailed understanding where and how different interactions occur. We are now at the point where we can study the reaction (phase change, star formation) of the multiphase ISM (molecular, atomic, ionized) to environmental interactions.

1 Introduction

During cosmic evolution, matter, which is sufficiently close to a massive object, decouples from the Hubble flow and is accreted by the massive object. Accretion takes place mostly within filamentary structures. In this way, a galaxy cluster gains mass through infall of dark matter, gas, and stars. Galaxy clusters have typical masses of 10^{14}–10^{15} $M_{\odot}$ and sizes of several Mpc. They contain hundreds of galaxies. The mass budget is dominated by dark matter which contributes about 80% to the total mass; galaxy clusters are thus the objects with the highest dark matter fraction in the universe. A tenuous ($\sim 10^{-3}$ cm^{-3}) gas sits in the gravitational potential of the whole cluster. When gas falls into the galaxy cluster, accretion shocks heat it to the Virial temperature (10^{7-8} K; Sarazin 1986). It then stays hot, because of the exceedingly long cooling time of such a low-density gas. This X-ray emitting cluster atmosphere represents typically 10–20% of the total cluster mass. The optically visible side of a galaxy cluster, i.e., the galaxies, only account for 5–10% of the cluster mass. The large-scale evolution of a galaxy cluster is thus mainly governed by gravitation and gas heating.

On smaller scales, things are more complex. Together with the dark matter and gas, galaxies fall into the cluster. These galaxies might have been isolated or assembled in groups before infall. Once they enter the cluster, their evolution can change radically through interactions with their environment. It is known for decades that galaxy populations in clusters are very different from those in the field, i.e., outside galaxy groups and clusters. The morphological type of a galaxy is closely related to its environment (Dressler 1980; Whitmore and Gilmore 1991): whereas about 80% of the field galaxies are spirals, this fraction drops to ~50% at the cluster outskirts and becomes almost zero in cluster cores.

Even within the disk galaxy population, cluster spiral galaxies are redder and have less star formation than field galaxy of similar Hubble types (Kennicutt 1983; Gavazzi et al. 2006a). Since morphology is closely related to the star formation history of a galaxy, it is not surprising that the average galaxy properties related to star formation also depend on local density (Hashimoto et al. 1998; Lewis et al. 2002; Gómez et al. 2003; Kauffmann et al. 2004; Balogh et al. 2004). The fraction of early-type galaxies and passive non-star-forming galaxies both grow with time during cluster evolution. The rate at which these fractions change depends sensitively on environment, i.e., the local density of galaxies (Dressler et al. 1997; Poggianti et al. 1999; Smith et al. 2005; Postman et al. 2005; Moran et al. 2007). Whereas in the cluster cores the early-type fraction has increased steadily with time from 70% at $z = 1$ to 90% at the present epoch, in intermediate-density regions corresponding to groups and the accretion regions of rich clusters, significant evolution appears to begin only after $z = 0.5$. At the same time, the fraction of blue star-forming galaxies in clusters decreases with time (Butcher–Oemler effect; Butcher and Oemler 1978, 1984).

The analysis of a spectroscopic catalog of galaxies in ten distant clusters (Dressler et al. 1999; Poggianti et al. 1999) has shown that the galaxy populations of these clusters are characterized by the presence of a large number of post-starburst galaxies. Poggianti et al. (1999) concluded that the most evident effect due to the cluster environment is the quenching of star formation rather than its enhancement. They found two different galaxy evolution timescales in clusters: (i) a rapid halt of star formation activity (~1 Gyr) and (ii) a slow transformation of morphology (several Gyr).

To understand the galaxy transformation with time, we have to understand when, where, and which interaction between a galaxy and its cluster environment changes the galaxy's aspect. The ideal place to find answers to these questions is a local cluster where the interaction mechanisms can be studied in detail at high spatial resolution (≤ 1 kpc) and where the whole range of galaxies to low luminosities can be observed. Therefore, this chapter is focused on the Virgo, Coma, and Abell 1367 clusters and to a lesser extent, on the Norma and Fornax clusters. At this point, it should be noted that galaxy clusters can be very different. Whereas the Virgo cluster is a dynamical young cluster with a high spiral fraction and a peaked X-ray emission (cooling core cluster), the Coma cluster is more relaxed and spiral-poor and has a much more extended X-ray emission distribution.

The consequences of a galaxy–cluster interaction can be observed in the integrated and resolved properties of a cluster galaxy. The seminal article on late-type galaxy evolution in local clusters of Boselli and Gavazzi (2006) is mainly based on the integrated properties of cluster galaxies at multiple wavelengths. The fundamental review on the relation between physical properties and environment of Blanton and Moustakas (2009) is based on statistical results from recent large surveys of nearby galaxies including the SDSS. Therefore, special attention will be payed to the resolved properties of Virgo spiral galaxies.

2 Galaxy Populations in Groups and Clusters

As many as 50–70% of all galaxies reside in groups of galaxies (Eke et al. 2005). These groups contain less than ~100 galaxies and have total masses of $\sim 10^{13}\ M_\odot$. Studies of groups at low redshift have revealed them to be a heterogeneous population with their galaxy populations varying from cluster-like to field-like (Zabludoff and Mulchaey 1998). Groups showing extended X-ray emission tend to have a significant fraction of early-type galaxies and a dominant early-type galaxy at the group center (Mulchaey and Zabludoff 1998; Mulchaey et al. 2003; Osmond and Ponman 2004; Jeltema et al. 2007). About 30% of galaxy groups are dominated by elliptical or lenticular galaxies (Croston et al. 2005).

The group environment changes the morphology of galaxies. McGee et al. (2008) compared the fractional bulge luminosities of galaxies in groups at $0.3 < z < 0.55$ to a similarly selected group sample at $0.05 < z < 0.12$. They found that, at both epochs, the group and field fractional bulge luminosity distributions differ significantly, with the dominant difference being a deficit of disk-dominated galaxies in the group samples. The group environment thus favors the formation of bulge-dominated early-type galaxies.

The same phenomenon is found in galaxy clusters. From his study of ~6,000 galaxies in 55 rich clusters, Dressler (1980) showed that the fraction of early-type galaxies increases with projected galaxy density: elliptical galaxies prevail in high-density regions, i.e., in the cores of massive clusters, whereas spiral galaxies are the dominant population in low-density regions. Lenticular galaxies occupy regions with intermediate to high galaxy densities. This is known as the morphology–density relation.

Moreover, the fraction of spiral galaxies in a galaxy cluster increases with increasing redshift (Dressler et al. 1997; Fasano et al. 2000). The spiral population in distant clusters consists of the great majority of blue galaxies responsible for the Butcher–Oemler effect, as well as a sizeable fraction of the red population (Dressler et al. 1999; Poggianti et al. 1999). Coupled to the increase in the spiral fraction, the S0 galaxies at intermediate redshifts are proportionately (two to three times) less abundant than in nearby clusters, while the fraction of ellipticals is already as large or larger (Dressler et al. 1997).

A morphology–density relation also exists for dwarf galaxies: dwarf elliptical galaxies are more frequent in dense environments, while dwarf irregulars are ubiquitous (Binggeli et al. 1990; Sabatini et al. 2005).

In galaxy clusters, a morphological segregation with respect to the distance from the cluster center is observed for $R < 0.5$ Mpc (Dressler 1980; Whitmore et al. 1993). Thomas and Katgert (2006) claimed that this morphology–radius relation is mainly due to the different radial distributions of bright elliptical and late-type galaxies.

Environment thus influences the morphological mix of galaxies in regions of different galaxy density and/or cluster radius. Is this mix already in place before the assembling of a massive structure as a galaxy cluster (nature) or is it established by galaxy evolution within a dense environment (nurture)? Environment does play an important role for the evolution of late-type spiral galaxies. In the following, it will be discussed how interactions between a galaxy and its environment can affect its ecology and ultimately its morphology.

3 Interaction Types

One can distinguish two different classes of interactions based on gravitation or gas physics. The first class includes galaxy–galaxy and galaxy–cluster tidal interactions. The second class involves the hot intracluster medium through which the galaxy is moving at a high speed. Whereas gravitational interactions act in the same way on all components of a galaxy (dark matter, stars, and gas), hydrodynamic interactions only affect the galaxy's interstellar matter.

3.1 Gravitational Interactions

The tidal interaction of a galaxy and the *gravitational potential of the whole cluster* is compressive within the cluster core. A disk parallel to the orbital plane develops a transient spiral pattern. If the disk is inclined with respect to the orbital plane, it is transiently compressed and an initially circular disk is deformed into an ellipse (Byrd and Valtonen 1990; Valluri 1993; Henriksen and Byrd 1996). The compression increases with decreasing distance to the cluster center.

Direct *galaxy–galaxy encounters* affect a spiral galaxy significantly if the impact parameter is of the order of or smaller than

$$b = r\left(2\frac{M}{m}\right), \tag{5.1}$$

where r and m are the size and mass of the galactic disk at the optical radius and M is the mass of the perturbing galaxy. In addition, the relative velocity between the galaxy has to be close to the rotation velocity of the spiral galaxies (100–200 km s^{-1}). The prominent example for a tidal interaction in the Virgo cluster is NGC 4438 (➲ *Fig. 5-1*).[1]

Fast encounters with relative velocities much larger than the rotation velocity lead to tidal shocks, which heat the systems. The increased kinetic energy causes the system to expand and cool. A prograde encounter where the orbital angular momentum is aligned with the rotational angular momentum leads to strong tidal arms, whereas a retrograde encounter gives rise to only mild tidal distortions. The seminal article of Struck (1999) gives a detailed description of tidal effects during galaxy–galaxy encounters.

Since the velocity of cluster galaxies is large (400–1,000 km s^{-1}), slow galaxy–galaxy encounters are rare once a galaxy has entered the cluster core. However, since a significant fraction of infalling galaxies are assembled in groups, slow encounters can occur at the periphery of clusters (preprocessing, Dressler 2004). On the other hand, multiple tidal shocks induced by rapid flybys of massive galaxies lead to an expansion of the stellar disk. This effect is termed *galaxy harassment* (Moore et al. 1996, 1999). The probability of close encounters with massive galaxies is higher in the cluster core, where the galaxy density is highest. The disk stars and gas loosened from the galaxy by multiple flybys are then stripped by the gravitational potential of the

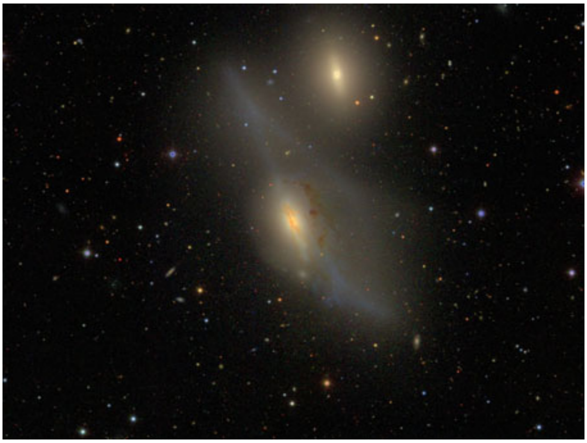


Fig. 5-1
SDSS *gri* image of the NGC 4438/4435 system in the Virgo cluster

[1] The strength parameter for a galaxy–galaxy collision has been defined by Gerber and Lamb (1994) as $S = 2GM/(bv_{\mathrm{rel}}v_{\mathrm{rot}})$, where v_{rel} is the relative velocity between the two galaxies and v_{rot} is the rotation velocity of the primary galaxy.

whole cluster. Since it is easier to strip matter from a galaxy with a shallow potential well, galaxy harassment most efficiently affects low-mass systems and might lead to a morphological transformation. However, in most cases harassment does not change the morphology of massive, high surface brightness disk galaxies (Moore et al. 1999).

Tidal shocking through multiple nearby high-speed encounters and the cluster potential depends on the galaxy orbit within the cluster. A highly eccentric orbit leads the galaxy close to the cluster center where both effects are strong.

3.2 Hydrodynamical Interactions

Whereas gravitational interactions act equally on the dark matter, stellar, and gaseous content of a spiral galaxy, hydrodynamical interactions take place between the hot intracluster gas and the interstellar medium of the galaxy. There is evidence that the star formation rate of a massive spiral galaxy of about 1 $M_\odot$/year has to be balanced by accretion of cold gas (Sancisi et al. 2008) and hot gas (Rasmussen et al. 2009) from an extended halo. In the absence of this gas halo, star formation is consuming the interstellar medium within a few Gyr. Larson et al. (1980) were the first to propose a halt of external gas accretion in cluster spiral galaxies. Bekki et al. (2002) developed this concept of *starvation* or *strangulation* by studying the stripping of an extended diffuse gas halo by ram pressure from the intracluster medium. They concluded that the diffuse halo is stripped if the galaxy orbit is eccentric, leading it to a pericenter distance three times larger than the cluster core radius. Once the halo is stripped, star formation will decrease and the galaxy becomes red.

A galaxy which enters a cluster on an eccentric orbit accelerates due to the gravitational potential of the cluster. At the same time, it encounters an intracluster medium of increasing density when it approaches the cluster center. From the point of view of the galaxy, which moves at a high speed through the cluster atmosphere, a wind due to ram pressure blows on its interstellar medium. If this ram pressure wind is higher than the gravitational restoring force due to the stellar disk, the gas is removed or stripped from the galactic disk (Gunn and Gott 1972):

$$\rho_{\mathrm{ICM}} v_{\mathrm{gal}}^2 > 2\pi G \Sigma_* \Sigma_{\mathrm{ISM}} \sim \frac{v_{\mathrm{rot}}^2}{R_{\mathrm{str}}} \Sigma_{\mathrm{ISM}}, \tag{5.2}$$

where G is the gravitation constant, ρ_{ICM} the intracluster medium density, v_{gal} the galaxy velocity with respect to the intracluster medium, $\Sigma_{*/\mathrm{ISM}}$ the stellar/gas surface density, v_{rot} the rotation velocity, and R_{str} the stripping radius. Gas which is located at $R > R_{\mathrm{str}}$ is removed via momentum transfer from the galactic disk by *ram pressure stripping*. The associated timescale is short t_{rps} ~10–100 Myr (Abadi et al. 1999; Vollmer et al. 2001a; Roediger and Brüggen 2006). One of the best examples of ongoing ram pressure stripping is the Virgo cluster spiral galaxy NGC 4522 (➋ *Fig. 5-2*).

Whereas momentum transfer ram pressure stripping only removes the gas of the outer disk, continuous gas removal from the whole gas disk proceeds via viscous stripping (Nulsen 1982). This stripping mode dominates if the galaxy moves edge-on through the cluster atmosphere. Viscosity can be classical for laminar flows or turbulent depending on the effective Reynolds number of the hot gas

$$Re = 2.8 \left(\frac{r_{\mathrm{gal}}}{\lambda_{\mathrm{ICM}}} \right) \left(\frac{v_{\mathrm{ISM}}}{c_{\mathrm{ICM}}} \right), \tag{5.3}$$

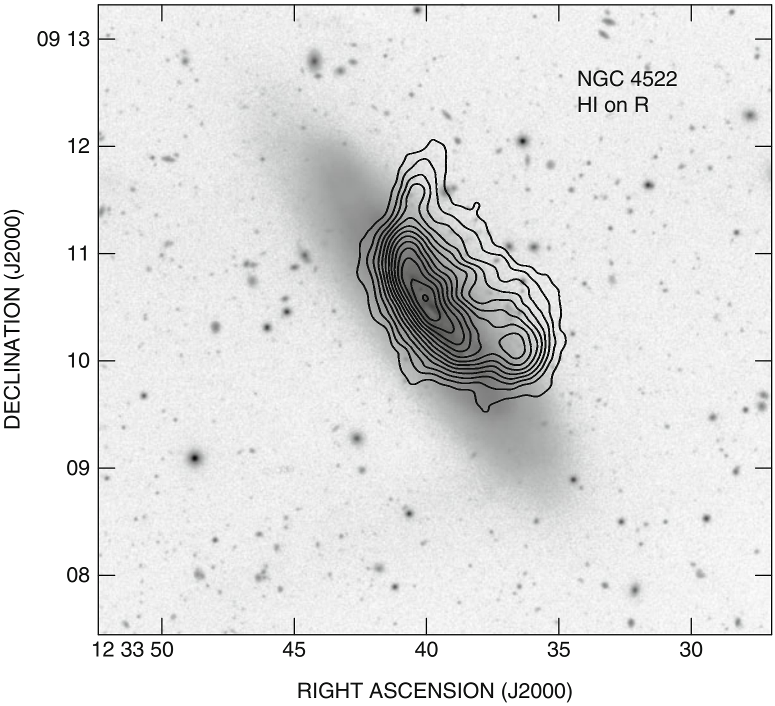


Fig. 5-2
The ram pressure stripped Virgo cluster spiral galaxy NGC 4522. *Gray scale*: stellar disk in the R-band image. *Contours*: HI surface density (From Kenney et al. 2004)

where λ_{ICM} is the characteristic length scale of the intracluster medium viscosity and c_{ICM} is the sound speed of the intracluster medium. If the characteristic length scale is the mean free path of ions in the intracluster medium,

$$\lambda_{\mathrm{ICM}} \sim 11 \left(\frac{T_{\mathrm{ICM}}}{10^8\ \mathrm{K}} \right) \left(\frac{10^{-3}\ \mathrm{cm}^{-3}}{n_{\mathrm{ICM}}} \right) \mathrm{kpc}, \tag{5.4}$$

where n_{ICM} is the number density of the intracluster medium. If $Re \geq 30$, the flow is expected to be turbulent. Turbulence is generated by Kelvin-Helmholtz instabilities at the interface between the intracluster and the interstellar medium. The gas disk which is not stripped by ram pressure is stable against Rayleigh-Taylor instabilities (Roediger and Hensler 2008). The mass loss rate in the two cases are

$$\dot{M}_{\mathrm{laminar}} = 4.3\pi r_{\mathrm{gal}} \rho_{\mathrm{ICM}} \lambda_{\mathrm{ICM}} c_{\mathrm{ICM}} \ \text{ and } \ \dot{M}_{\mathrm{turb}} = \pi r_{\mathrm{gal}}^2 \rho_{\mathrm{ICM}} v_{\mathrm{gal}}. \tag{5.5}$$

Both gas stripping rates are similar. The timescale of turbulent viscous stripping is

$$\tau_{\mathrm{visc}} = \frac{\dot{M}_{\mathrm{turb}}}{M_{\mathrm{ISM}}} \sim \frac{\rho_{\mathrm{ISM}} H}{\rho_{\mathrm{ICM}} v_{\mathrm{gal}}} \sim \frac{\Omega v_{\mathrm{turb}}}{\pi G \rho_{\mathrm{ICM}} v_{\mathrm{gal}}}, \tag{5.6}$$

where the ISM density is $\rho_{\text{ISM}} \sim \Omega^2/(\pi G)$ (Vollmer and Beckert 2002), Ω the angular velocity of the ISM, H the ISM disk height, and v_{turb} the velocity dispersion of the ISM. The critical ram pressure for turbulent viscous stripping is

$$\rho_{\text{ICM}} v_{\text{gal}}^2 > \rho_{\text{ISM}} \frac{\lambda_{\text{KH}}}{2\pi} \frac{v_{\text{rot}}^2}{R_{\text{str}}} = \frac{\lambda_{\text{KH}}}{2\pi H} \frac{v_{\text{rot}}^2}{R_{\text{str}}} \Sigma_{\text{ISM}}, \tag{5.7}$$

where λ_{KH} is the dominant wavelength for the gas ablation by Kelvin-Helmholtz instability (see, e.g., Mori and Burkert 2000). For $\lambda_{\text{KH}} \sim 6\,H$, this criterion is similar to the classical Gunn and Gott criterion (➋ 5.2). The stripping radius for viscous edge-on stripping is somewhat larger but comparable to that of momentum transfer face-on stripping (Marcolini et al. 2003). The instabilities lead to mixing of the intracluster medium into the ISM, which leads to a decrease of the ISM density. For $n_{\text{ICM}} = 10^{-4}$ cm^{-3}, $\Omega = 10^{-8}$ year^{-1}, $v_{\text{turb}} = 10$ km s^{-1}, and $v_{\text{gal}} = 1{,}000$ km s^{-1}, the viscous stripping timescale is $\tau_{\text{visc}} \sim 3$ Gyr, much longer than that of classical momentum transfer ram pressure stripping (10–100 Myr). The magnetic fields contained in the ISM and intracluster medium might suppress Kelvin-Helmholtz instabilities (Landau and Lifshitz 1960; Spitzer 1978) and stripping might become laminar. Moreover, if the fields are stretched along the surface, thermal conduction will be suppressed (Vikhlinin et al. 2001).

4 Simulations

Since the work of Toomre and Toomre (1972), gravitational interactions between two galaxies have been simulated extensively (see, e.g., Combes et al. 1988; Barnes and Hernquist 1996; Duc et al. 2004). Today's simulations of gravitational interactions include the interstellar medium of the galaxies and a recipe for star formation. Byrd and Valtonen (1990) and Valluri (1993) investigated gravitational interactions between a disk galaxy and the gravitational potential of the whole cluster. Since the impact of this interaction on a galaxy is small, this subject did not receive further attention.

In the recent past, most progress has been made on the modeling of ram pressure stripping events. Therefore, this section is focused on the simulations of the hydrodynamic interaction between the intracluster medium and the ISM. The ISM can be simulated either by a continuous description (Eulerian hydrodynamics in 2D or 3D), a discrete description (sticky particles), or a discrete-continuous hybrid description (smoothed particles hydrodynamics, SPH). The common point of all methods is that the gas is treated as a collisional phase, whereas dark matter and stars are collisionless. The discrete description has no shocks nor instabilities and a finite penetration length of the intracluster medium into the ISM. This penetration length is 0 by definition in SPH. The latter method handles shocks and Rayleigh-Taylor instabilities, but it is not able to produce Kelvin-Helmholtz instabilities. The continuous description solves Euler's equations for compressible gas dynamics. Either the gas is assumed to be adiabatic or gas cooling is explicitly implemented. Modern codes contain an adaptive mesh refinement to resolve small-scale structures as shocks (see, e.g., Fryxell et al. 2000). Since the ISM is neither a continuous medium nor exclusively made of clouds, one has to chose the numerical method which is well adapted for the investigated astrophysical problem.

4.1 Spherical Galaxies

The bulk of gas in elliptical galaxy is hot, has a high volume filling factor, and can be detected in X-rays. It has to be modeled by a continuous prescription. The gas is supported against gravity by thermal pressure and rotation is negligible. Complete ram pressure stripping from a spherical galaxy requires that ram pressure exceeds the thermal pressure at the center of the gravitational potential well of the galaxy. For incomplete stripping, hydrodynamical simulations show a leading bow shock and a weak gravitationally focused wake or tail behind the galaxy.

Takeda et al. (1984), Gaetz et al. (1987), and Balsara et al. (1994) studied the influence of constant ram pressure in 2D. The two latter simulations included gas cooling and replenishment by stellar mass loss. Gaetz et al. (1987) determined several important parameters affecting the stripping efficiency. For not too high galaxy velocities, Balsara et al. (1994) found that the galaxy accretes mass from the downstream side into the core. Stevens et al. (1999) continued the work of Balsara performing a parameter study, for parameters appropriate for different clusters, ranging from cool clusters or groups up to hot clusters. They provide synthetic flux and hardness maps, as well as surface brightness profiles which can be directly compared to observations. Mori and Burkert (2000) studied ram pressure stripping of a cluster dwarf spheroidal galaxy without cooling and gas replenishment. They found that the gas in dwarf galaxies is rapidly removed in a typical cluster environment by ram pressure stripping. Stripping proceeds in two phases: (i) instantaneous ram pressure stripping via momentum transfer and (ii) gas ablation by Kelvin-Helmholtz instabilities with a timescale of ~1 Gyr.

4.2 Disk Galaxies

The ISM in disk galaxies is supported by rotation. For a face-on ISM-intracluster medium interaction, gas stripping occurs mainly via rapid momentum transfer. For edge-on interactions, slow viscous stripping or gas ablation by Kelvin-Helmholtz instabilities is responsible for ISM removal. For intermediate angles between the disk plane and the direction of the ram pressure wind, both mechanisms are at work. Since SPH and sticky particle codes do not produce Kelvin-Helmholtz instabilities, they cannot simulate turbulent viscous stripping.

All simulations validate the Gunn and Gott formula (➋ 5.2) and yield similar short timescales for ram pressure stripping via momentum transfer (10–100 Myr; Jáchym et al. 2007, 2009). Whereas sticky particle and SPH simulation produce similar results, Eulerian hydrodynamical simulations show differences in (i) the dependence between the stripped mass fraction and inclination angle between the disk plane and ram pressure wind and (ii) gas back-fall after the ram pressure peak. In the following, an overview is given on simulations using the different numerical techniques.

Tosa (1994) studied the influence of ram pressure on a disk galaxy treating the gas as test particles. Vollmer et al. (2001a) used a sticky particle code including the effects of ram pressure stripping to study the influence of (i) time-dependent ram pressure due to the galaxy orbit within the cluster and (ii) the inclination angle between the orbital and the disk plane on gas removal. They showed that ram pressure can lead to a temporary increase of the central gas surface density. In some cases, a considerable part of the total atomic gas mass (several 10^8 $M_\odot$) can fall back onto the galactic disk after the stripping event. A quantitative relation between the orbit parameters and the resulting gas loss was derived containing explicitly the inclination angle between the orbital and the disk plane. Jáchym et al. (2009) proposed an alternative and physically motivated relation depending on the disk tilt angle and on the column density of the encountered intracluster medium.

Abadi et al. (1999) modeled a constant ram pressure wind on a disk galaxy using SPH for different values of ram pressure and different inclination angles between the orbital and the disk plane. Schulz and Struck (2001) used an SPH code allowing for gas cooling. They observed a dependency of the stripped gas mass on the inclination angle. Not all the gas stripped from the disk escapes immediately from the halo, some gas can linger for times of order 100 Myr. Under the action of ram pressure, the gas disk is displaced relative to the halo center and compressed. In this way, the formation of numerous flocculent spirals is triggered. These waves transport angular momentum outward, resulting in further compression of the inner disk. This "annealing" process makes the inner disk, which contains much of the total gas mass, resistant to further stripping. Jáchym et al. (2007, 2009) investigated the evolution of the ISM which undergoes time-dependent ram pressure at various inclination angles with SPH. They highlighted the significant dependence of the stripping efficiency on the duration of a short ram pressure pulse. Their simulations confirmed the general trend of less stripping at orientations close to edge-on. The dependence on the disk tilt angle is more pronounced for compact intracluster medium distributions; however it almost vanishes for strong ram pressure pulses. The role of ram pressure stripping on star formation was studied by Kronberger et al. (2008).

Eulerian hydrodynamical simulations of ram pressure stripping of disk galaxies became possible only in the last decade in 2D (Roediger and Hensler 2005) and 3D (Quilis et al. 2000; Roediger and Brüggen 2006; Tonnesen and Bryan 2009). Marcolini et al. (2003) addressed the role of ram pressure stripping for dwarf disk galaxies. Roediger and Brüggen (2006) investigated the role of inclination between the orbital and the disk plane for constant ram pressure (➲ *Fig. 5-3*). They found that the inclination angle does not play a major role for the mass loss as long as the galaxy is not moving close to edge-on ($i > 30^\circ$). This different behavior compared to sticky particles and SPH is probably due to the action of strong Kelvin-Helmholtz instabilities. The stripping of disk galaxies on realistic orbits within the cluster was studied by Roediger and Brüggen (2007). These authors observed a repeated backfall of stripped gas prior to pericenter passage. In contrast to SPH and sticky particle simulations, there is no backfall after pericenter passage. Tonnesen and Bryan (2009) included radiative cooling in their high-resolution simulations to model the reaction of clumpy ISM to ram pressure stripping. In their simulations, lower density gas is stripped quickly from any radius of the galaxy, and the higher density gas can then be ablated via Kelvin-Helmholtz instabilities. The overall stripping timescale is thus longer than in previous simulations. The stripped gas tails are studied in Roediger et al. (2006), Roediger and Brüggen (2008), Tonnesen and Bryan (2010), and Tonnesen et al. (2011).

The combined influence of tidal interactions and ram pressure stripping was addressed by Vollmer (2003), Vollmer et al. (2005a, b), and Kapferer et al. (2008). Tidal interactions can push the outer ISM to larger galactocentric radii. In addition, the ISM has often a lower surface density. This tidally loosened gas is easily stripped by ram pressure. Tidal interactions thus increase the stripping efficiency.

Three-dimensional Eulerian hydrodynamic simulations of a whole galaxy cluster are presented in Domainko et al. (2006), Kapferer et al. (2007), and Tonnesen et al. (2007). These simulations generally show huge gas tails behind the stripped galaxies. In these simulations, the intracluster medium is not static but moving. Depending on the galaxy's velocity vector $\vec{v}_{\rm ICM}$, this changes ram pressure from $\rho_{\rm ICM} v_{\rm gal}^2$ to $\rho_{\rm ICM} |\vec{v}_{\rm gal} - \vec{v}_{\rm ICM}|^2$. This can have significant consequences for a galaxy moving against the ram pressure wind (see NGC 4522 in ➲ Sect. 8.3).

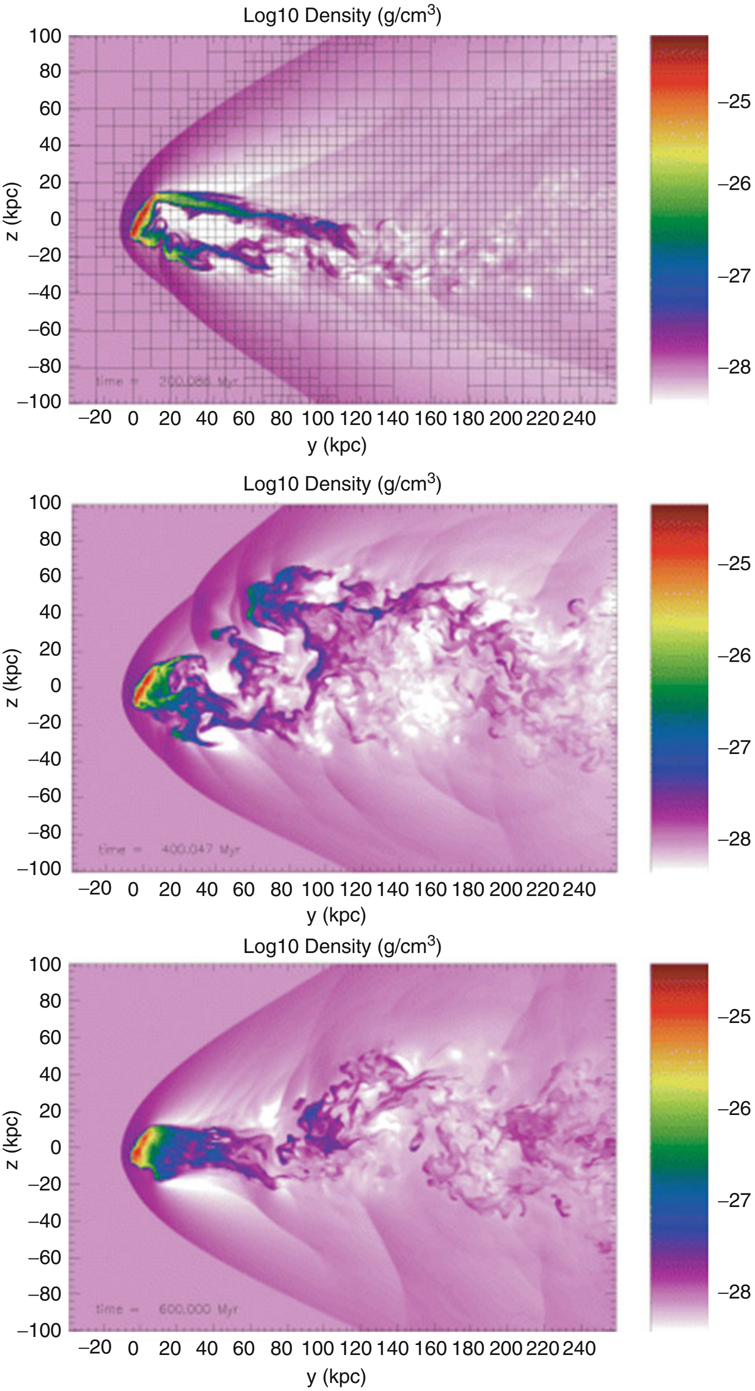


Fig. 5-3

Three-dimensional Eulerian hydrodynamic simulation of a ram pressure stripping event. The local gas density is shown color-coded for different time-steps: *top* figure corresponds to a time of 200 Myr, the *middle* one to 400 Myr, and the *bottom* one to 600 Myr (From Roediger et al. 2006)

5 The Multiphase ISM

In a widely accepted picture (see, e.g., Kulkarni and Heiles 1988; Spitzer 1990; McKee 1995), the ISM consists of five different phases which are listed in *Table 5-1*. About 80% of the total gas mass is neutral and 50% is in form of clouds or filaments. The warm/cold neutral phase is observable in Hɪ line emission/absorption and the molecular phase in CO line emission. Hα emission traces the dense warm ionized medium (diffuse gas and Hɪɪ regions) and X-ray emission the hot ionized gas phase. The CO emission in spiral galaxies is concentrated in the inner disk where giant molecular clouds are ubiquitous. In field spiral galaxies, Hα emission is detectable over the whole stellar disk. The extent of Hɪ emission is usually 1.5–2 times larger than the galaxy's optical diameter (R_{25}). Cosmic ray electrons, which are accelerated in supernova shock waves to relativistic velocities, are observable in the radio continuum at wavelengths ≥6 cm. The associated magnetic field is traced by the polarized radio continuum emission. In the following, an overview is given on the observations of the different ISM phases in groups and clusters with emphasis on nearby clusters.

5.1 Atomic Hydrogen

The ISM at galactic radii larger than ~2 scale-lengths of the optical disk of a spiral galaxy is mostly made of atomic hydrogen. The Hɪ disk of isolated galaxies typically extends up to 1.5–1.8 optical radii (Cayatte et al. 1990; Salpeter and Hoffman 1996; Broeils and Rhee 1997; Walter et al. 2008). Since the outer gas of a spiral galaxies is less strongly bound by the gravitational potential, it is most vulnerable against external perturbations. Therefore, the Hɪ emission of a cluster and group spiral galaxy is the most sensitive tracer of environmental interactions.

Hɪ-deficient galaxies are observed in group environments. This indicates that the environment affects the gas content of the group galaxies. Since the galaxy velocity dispersion and the intragroup gas density are small, slow tidal interactions are the main driver of galaxy evolution within groups. The fraction of Hɪ-deficient galaxies depends on the evolutionary stage of the group. Evolved groups display an X-ray halo and a high fraction of early-type galaxies. Kilborn et al. (2009) determined the Hɪ deficiency of galaxies within 16 galaxy groups. They found that around two-thirds of Hɪ-deficient galaxies were preferentially located at a

Table 5-1
The properties of the different phases of the ISM in spiral galaxies (From Boulares and Cox 1990)

	T (K)	n (cm^{-3})	$<n>$ (cm^{-3})	v_{turb} (km s^{-1})	Φ_V	H (pc)	M/M_{tot} (%)
Hot ionized	~10^6	..	0.002	..	0.5	3,000	4
Warm ionized	~8,000	0.3–10	0.025	~10	0.2	900	14
Warm neutral	~8,000	0.1–10	0.1	~10	0.3	400	31
Cold neutral	~100	10–1,000	0.3	~6	0.02	140	25
Molecular	~10	>100	0.6	~6	0.001	70	26

T temperature, n: density, < n > mean density, v_{turb} dispersion velocity, Φ_V volume filling factor, H scale height, M/M_{tot} percentage of the total gas mass projected distance of less than 1 Mpc from the group center.[2] Sengupta et al. (2007) observed 13 galaxies from four X-ray bright groups having disturbed and often truncated HI disks. Kern et al. (2008) imaged 16 HI sources in six galaxy groups. They found several interacting systems with extended HI envelops. The Hickson compact groups (HCGs), owing to their high number density coupled with low-velocity dispersions, undergo frequent tidal interactions, distortions, and mergers between group members (Hickson et al. 1992; Mendes de Oliveira and Hickson 1994). Huchtmeier (1997) determined the HI content of 54 Hickson Compact groups. Many of these groups show extreme HI deficiencies as compared to nearby groups. In their sample of 72 Hickson compact groups, Verdes-Montenegro (2001) found that galaxies are, on average, deficient in HI by a factor of ~2 compared to loose groups. Tidal interactions between galaxies in groups thus can efficiently remove atomic gas from galactic disks.

Davies and Lewis (1973) were the first who noticed a difference in the HI content of Virgo spiral galaxies compared to field spiral galaxies of the same morphological type. A first quantitative comparison sample for isolated galaxies was established by Haynes and Giovanelli (1984). Similar to Chamaraux et al. (1980), these authors introduced the HI deficiency:

$$def_{\mathrm{HI}} = \log \frac{M_{\mathrm{HI}}^{\mathrm{exp}}}{M_{\mathrm{HI}}^{\mathrm{obs}}}, \tag{5.8}$$

where $M_{\mathrm{HI}}^{\mathrm{obs}}$ is the observed and $M_{\mathrm{HI}}^{\mathrm{exp}}$ the expected HI mass of a galaxy of the same size and morphological type. A galaxy with an HI deficiency of $def_{\mathrm{HI}} = 1$ has thus lost 90% of its atomic hydrogen. The typical uncertainty of the HI deficiency is 0.25. Based on their comparison sample, Giovanelli and Haynes (1985) showed that a significant fraction of galaxies in dense clusters are HI deficient and that the HI deficiency increases with decreasing distance to the cluster center (➲ *Fig. 5-4*). Gavazzi (1987, 1989) and Gavazzi et al. (2005, 2006b) completed this picture for the Virgo cluster and the Coma supercluster region. The HI mass function of the Virgo Cluster differs significantly from that in the field, due to the combined effect of morphology segregation and the presence of HI-deficient objects (Gavazzi et al. 2005).

Solanes et al. (2001) continued the work of Giovanelli and Haynes (1985) by extending the sample to 1,900 galaxies within 18 clusters. They did not find a correlation between the fraction of HI-deficient spirals in a cluster and the clusters' global properties. The HI deficiency is related to the morphology of the galaxies and not to their optical size. The radial extent of the region with significant gas removal from galaxies can reach up to two Abell radii. Within this region, the fraction of HI-deficient spiral galaxies increases continuously toward the cluster center.

A blind Arecibo HI survey of the Virgo cluster (80 deg^2) showed that HI-rich Virgo galaxies are structurally similar to ordinary late-type galaxies. This is consistent with a scenario where a structural change of a galaxies implies a significant loss of atomic hydrogen. Moreover, less than 1% of early-type galaxies contain neutral hydrogen (>10^8 $M_\odot$; Gavazzi et al. 2008).

Imaging observations of the atomic gas content of cluster spiral galaxies showed that HI-deficient galaxies have truncated gas disks (Virgo cluster: Warmels 1988; Cayatte et al. 1990; Chung et al. 2009 (➲ *Fig. 5-5*); Coma cluster: Bravo-Alfaro et al. 2000). Galaxies near the cluster core (≤0.5 Mpc) have HI disks that are significantly smaller compared to their stellar disks. The most natural explanation for these findings is ram pressure stripping. Cayatte et al. (1994) divided their galaxy sample into four groups: (i) HI-normal galaxies, (ii) galaxies with mildly truncated HI and a central surface density which is lower than that of normal galaxies, (iii) galaxies with strongly truncated HI and a low central surface density, and (iv) anemic galaxies

[2] Galaxy groups have sizes of ~1 Mpc. The typical size of compact groups is about half this value.

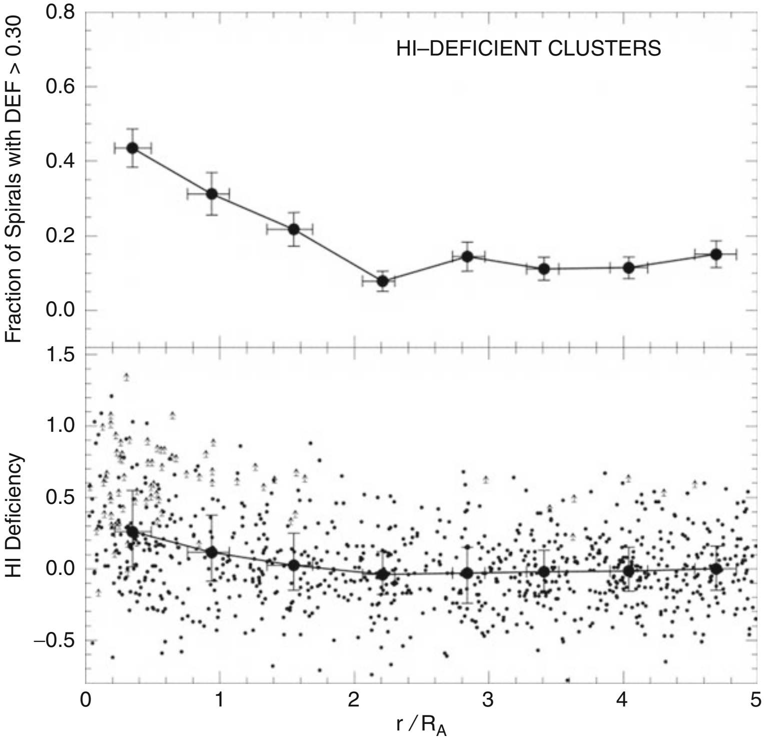


Fig. 5-4

***Top*: fraction of Hɪ-deficient galaxies in bins of projected radius from the cluster center for the superposition Hɪ-deficient clusters. Vertical error bars correspond to 1σ confidence Poisson intervals. The abscissae show medians and quartile values of the bins in radial distance. *Bottom*: Same as top panel for the measured Hɪ deficiency. Displayed are the medians and quartiles of the binned number distributions in Hɪ deficiency. *Small dots* show the radial variation of Hɪ deficiency for individual galaxies, while the *arrows* identify non-detections plotted at their estimated lower limits (From Solanes et al. 2001)**

which a mildly truncated and show an overall very low Hɪ surface density. In their sample of ~50 Virgo galaxies, Chung et al. (2007) found seven spiral galaxies with long Hɪ tails. These galaxies are found in intermediate- to low-density regions (0.6–1 Mpc in projection from M 87). The tails are all pointing roughly away from M 87, suggesting that these tails may have been created by a global cluster mechanism. A rough estimate suggests that simple ram pressure stripping could have indeed formed the tails in all but two cases. It should be noted that in three of the seven systems, a gravitational galaxy–galaxy interaction is involved.

5.2 Molecular Hydrogen

The ISM becomes mainly molecular in the inner part of the galactic disk. The H_2 surface density has about the same scale-length as the stellar disk (e.g., Leroy et al. 2008). The molecular

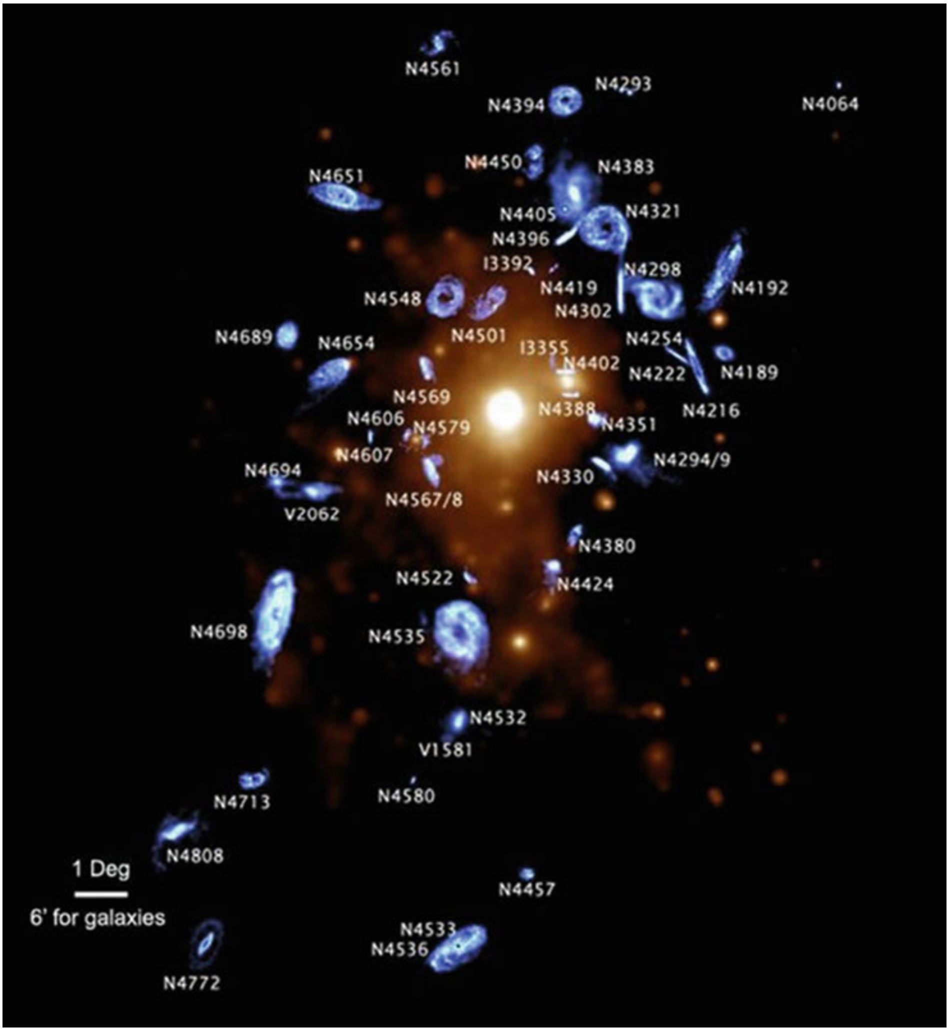


Fig. 5-5

Composite image of the HI distribution of the individual galaxies (in *blue*) overlaid on the ROSAT X-ray image (*orange*) by Böhringer et al. (1994). The galaxies are located at the proper position in the cluster, but each HI image is magnified by a factor 10 (From Chung et al. 2009)

gas is sitting deeply in the gravitational potential of the galaxy and is thus very hard to move or remove. However, in the absence of gas accretion, the molecular gas will be consumed by star formation within a few Gyr (Bigiel et al. 2008). Molecular hydrogen can only be indirectly observed via CO line emission. To derive H_2 column densities and masses, a conversion factor $X = N(H_2)/I(CO)$ has to be assumed. For massive galaxies, the standard value $X = 2 \times 10^{20}$ cm^{-2}(K km s^{-1})$^{-1}$ applies (Dame et al. 2001). For low-mass, low-metallicity systems, the conversion factor is substantially higher. Boselli et al. (2002) give a correlation between

X and the NIR H band luminosity of a galaxy. As the HI deficiency, the H_2 deficiency is defined as the logarithm of the ratio between the expected and measured H_2 mass of a galaxy.

Observations of the molecular gas in group galaxies are rare. Boselli et al. (1996) did not find any difference between the molecular gas content of spiral galaxies in Hickson Compact groups compared to isolated spirals. On the other hand, Leon et al. (1998) argued for an enhanced molecular gas content in compact group galaxies, especially for the most compact groups, suggesting that tidal interactions can drive the gas component inward, and concentrating it in the dense central regions, where it is easily detected.

CO observations of galaxies in the Virgo (Stark et al. 1986; Kenney and Young 1986; Boselli et al. 1995, 2002) and Coma cluster (Casoli et al. 1991, 1996) did not show an H_2 deficiency in these clusters, i.e., cluster spirals are indistinguishable from field spirals with respect to their molecular gas content. However, Fumagalli and Gavazzi (2008) found a weak correlation between the H_2 mass divided by the stellar mass and the HI deficiency.

5.3 Dust and Metallicity

Heavy elements are produced in massive stars and released by stellar winds and supernova explosions. Metallicity is thus closely linked to the star formation history of a galaxy. The spectrum of a stellar population does not only depend on its age but also on its metallicity. Little is known about environmental effects on metallicity.

Early-type galaxies with low-velocity dispersions in Hickson Compact groups show an enhanced [Mg/Fe] ratio and depleted metallicity [Z/H] with respect to their counterparts in the field (de la Rosa et al. 2007). This anomalous behavior is interpreted as evidence for the action of a mechanism that truncates star formation. This is expected after a merger event between two spiral galaxies.

The influence of environment on metallicity is still an open question. Skillman et al. (1996) determined the metallicity gradients of 9 Virgo spiral galaxies. They found that HI-deficient Virgo galaxies have larger mean abundances than field galaxies of comparable luminosity or Hubble type, while the spirals at the periphery of the cluster are indistinguishable from the field galaxies. There is also weak evidence of shallower abundance gradients in the HI-deficient Virgo spirals compared to the spirals on the periphery of the cluster. On the other hand, by reanalyzing the Skillman sample, Pilyugin et al. (2002) found that all Virgo periphery and core spirals have counterparts among field spirals. They concluded that if there is a difference in the abundance properties of the Virgo and field spirals, this difference appears to be small and masked by the observational errors. Lee et al. (2003) compared oxygen abundances of Virgo to local dwarf elliptical galaxies with the conclusion that there is no systematic difference between the two populations. Boselli and Gavazzi (2006) presented evidence that HI-deficient galaxies have on average larger metallicities than similar objects with a normal gas content.

Metals are injected into the interstellar medium by AGB stars, stellar winds, and supernova explosions. These metals aggregate to form dust which absorbs the energy of the stellar UV emission and reemits it in the infrared. The ISM dust-to-gas column density ratio is related to the metallicity. Molecular hydrogen forms on the surface of dust grains. The H_2 formation timescale thus depends on the dust-to-gas ratio or metallicity of the ISM (Tielens and Hollenbach 1985). In normal galaxies, the bulk of dust mass has temperatures around 20 K and is detected at wavelengths >100 μm. The bulk of the radiation is emitted at wavelengths between 60 and 200 μm from warm dust (~30 K). Dust is mixed with the ISM and therefore follows its distribution. Gas removal thus always implies dust removal. Bicay and Giovanelli (1987) derived far-infrared (IRAS 60–100 μm) properties for a sample of over 200 galaxies in seven clusters. Irrespective of the HI deficiency of a cluster galaxy, the sample consists almost entirely of infrared-normal galaxies. However, they found a lack of high luminosity ($L_{\rm IR} > 10^{11}$ L$_\odot$) compared to the field. The Herschel (100–500 μm) Virgo cluster luminosity functions show the same lack of very luminous galaxies. In addition, they do not have the large numbers of faint galaxies seen previously in surveys covering less dense environments (Davies et al. 2010). Popescu et al. (2002) derived temperatures and masses of cold dust from the IRAS and ISOPHOT 170 μm data. Using the same data, Boselli and Gavazzi (2006) argued that there is a tentative trend of smaller dust masses in galaxies of higher HI deficiency. This is expected if the dust which is associated to the ISM is stripped during a ram pressure stripping event. Based on Herschel data, Cortése et al. (2010b) confirmed this scenario by showing that galaxies with truncated HI disks also have truncated dust disks. Stickel et al. (2003) presented the first tentative detection of an intergalactic dust cloud in the region between galaxies near the stripped Virgo spiral galaxy, NGC 4402. The clearest evidence of dust displaced together with the ISM from the disk of the Virgo cluster spiral galaxy NGC 4438 is presented by Cortése et al. (2010a).

5.4 Cosmic Ray Gas

Electrons are accelerated to relativistic velocities in supernova shocks. In the presence of a magnetic field, they emit synchrotron radiation. The hot electrons in HII regions emit thermal bremsstrahlung. The radio continuum emission of non-starburst galaxies at wavelengths >6 cm is dominated by synchrotron emission whose emissivity is proportional to the density of relativistic electrons and the magnetic field strength to the power of ~2. In spiral galaxies, the radio continuum emission is closely related to the star formation rate and the far-infrared emission (Helou et al. 1985; Niklas et al. 1997; Niklas 1997). It is well possible that the influence of the cluster environment on the radio continuum emission of cluster galaxies depends on the cluster properties, especially on those of the intracluster medium. Jaffe and Gavazzi (1986), Andersen and Owen (1995), and Rengarajan et al. (1997) found an enhanced radio to far-infrared ratio in galaxies which are located in the cores of dense clusters. Gavazzi and Boselli (1999a) studied the radio luminosity function of Virgo cluster galaxies for early- and late-type galaxies separately. They found that late-type galaxies develop radio sources with a probability proportional to their optical luminosity, independently of their detailed Hubble type. In a second article, Gavazzi and Boselli (1999b) compared the radio luminosity functions of galaxies in different clusters to those of isolated galaxies. They concluded that the radio luminosity function of Virgo cluster galaxies is consistent with that of isolated galaxies, whereas the Coma cluster galaxies show an excess of radio emissivity. Moreover, Gavazzi and Boselli (1999b) suggested that the radio excess observed in dense cluster galaxies is probably due to ram pressure compression of the ISM and its associated magnetic field or shock-induced re-acceleration of relativistic electrons as proposed by Völk and Xu (1994). It should be kept in mind that in the absence of imaging observations, a radio excess can also be due to nuclear activity. Vollmer et al. (2004a) determined the spectral index of the radio continuum emission of 81 Virgo galaxies. They noted that galaxies showing flat radio spectra also host active centers. No clear trend appeared between the spectral index and the galaxy's distance to the cluster center.

6 Star Formation

As reviewed in ➋ Chap. 3, the star-formation activity of a galaxy can be determined using various tracers (Kennicutt 1998a): (i) dense HII regions around massive (>8 $M_\odot$) stars give rise to Hα line emission, (ii) other recombination and forbidden nebular lines are emitted by gas around massive stars, (iii) intermediate-mass (2–5 $M_\odot$) stars show strong UV emission, (iv) dust is heated by this UV emission and radiates at wavelengths between 15 and 60 μm, and (v) supernova shocks/remnants emit in the radio continuum. Since star formation approximatelly follows the ISM surface density via the Schmidt law (Schmidt 1963; Kennicutt 1998b), gas removal leads to a decrease of star formation in a group or cluster galaxy. On the other hand, gravitational interactions can lead to important tidal torques which compress the gas and lead to a temporarily enhanced star formation rate.

Wilman et al. (2005) showed that the fraction of galaxies with [O II] emission, a measure of star formation, is much higher in group galaxies at intermediate redshift ($z \sim 0.4$) than in the local universe. However, the group galaxies still exhibit suppressed star formation relative to the field at the same epoch. Although observations have found abundant traces of tidal interactions in Hickson Compact Group (HCG) galaxies (Mendes de Oliveira and Hickson 1994), their star formation levels are surprisingly similar to those found in isolated galaxies (e.g., Zepf and Whitmore 1991; Moles et al. 1994; Allam et al. 1996; Verdes-Montenegro et al. 1998; Iglesias-Páramo and Vílchez 1999). Johnson et al. (2007) looked at the Spitzer IRAC (3.6–8.0 μm) color space distribution of HCGs and found that the mid-infrared (MIR) colors of galaxies in HI gas-rich HCGs are dominated by star formation, while the MIR colors of galaxies in HI gas-poor HCGs are dominated primarily by stellar photospheric emission. Galaxies in the most gas-rich groups tend to be the most actively star forming. Galaxies in the most gas-poor groups tend to be tightly clustered around a narrow range in colors consistent with the integrated light from a normal stellar population. These authors infer an evolutionary sequence in which galaxies in gas-rich groups experience star formation and/or nuclear actively until their neutral gas is consumed, stripped, or ionized.

Kennicutt (1983) was the first to compare the star formation rate of cluster spirals to that of field spirals of similar Hubble type. He found that Virgo cluster spiral galaxies have on average lower star formation rates than their isolated counterparts. Based on Hα observations of 273 galaxies in the Virgo, Coma, Abell 1367, and Cancer clusters, Gavazzi et al. (2006a) concluded that, within each Hubble-type class, galaxies with normal HI content have twice the Hα equivalent width of their HI-deficient counterparts. The star formation rate per unit mass of high-luminosity spirals that are projected within one virial radius is about a factor of 2 lower than at larger clustercentric projected distances, whereas low-luminosity objects have similar Hα properties at all clustercentric radii.

The analysis of the Hα equivalent width of large galaxy samples with redshifts $0.03 \leq z \leq 0.1$ from the SDSS and the 2dF survey (Lewis et al. 2002; Gómez et al. 2003; Tanaka et al. 2004) showed that the overall distribution of star formation rates is shifted to lower values in dense environments compared with the field population. This is consistent with earlier findings by Hashimoto et al. (1998). The distribution of the star formation rate as a function of projected galaxy surface density shows a discontinuity or a break at a galaxy number density of ~1 Mpc^{-2}. This corresponds to a cluster radius of about 4 Mpc or ~3 Virial radii. In the Tanaka et al. (2004) sample, only faint galaxies show this break. The morphology of these galaxies also changes significantly. It seems thus difficult to disentangle the star formation-density relation from the morphology–density relation. It should be noted that all galaxies used in these samples are of high luminosity according to the definition of Gavazzi et al. (2006a).

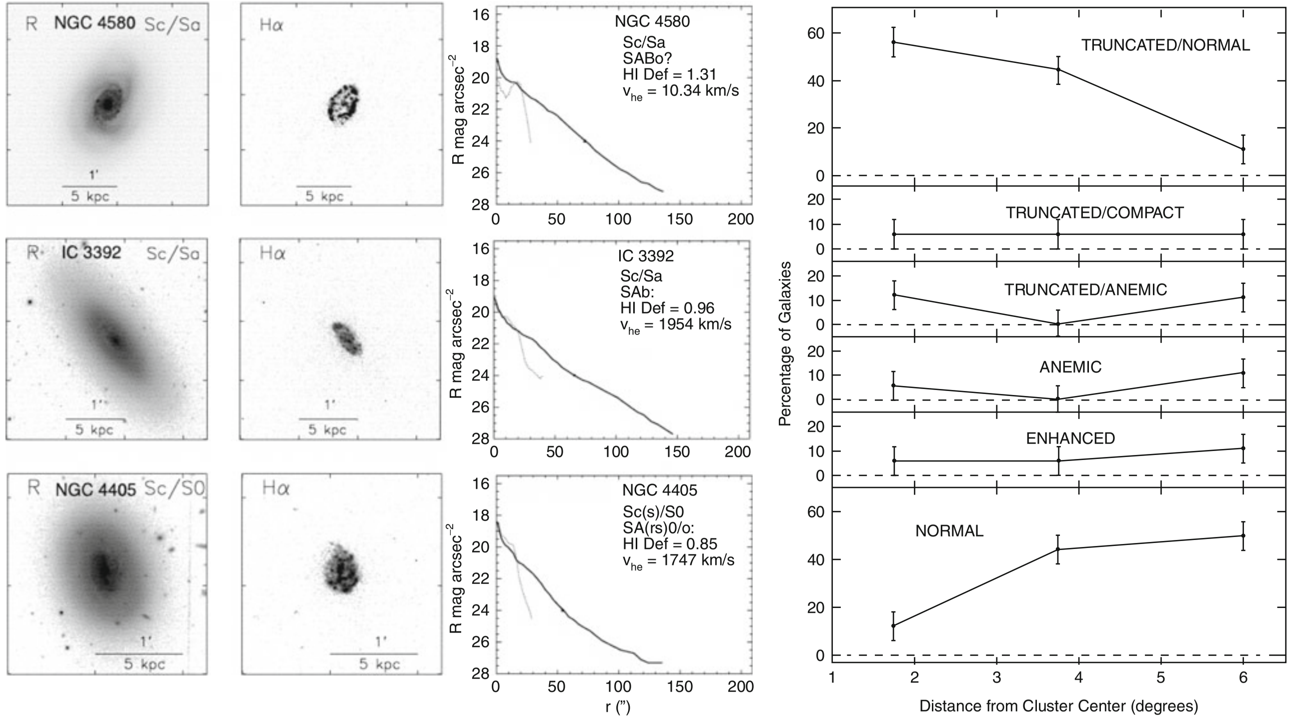
Imaging observations of the star formation of larger samples of cluster galaxies are rare. Koopmann and Kenney (2004a, b) divided the Hα morphology of 52 Virgo spiral galaxies into several categories: (i) normal (37% of the sample), anemic (6%), enhanced (6%), and (spatially) truncated (52%). Anemic galaxies have a significant lower overall Hα surface brightness. Truncated galaxies are further subdivided on the basis of their inner star formation rates into truncated/normal (37%), truncated/compact (6%), truncated/anemic (8%), and truncated/enhanced (2%). The fraction of anemic galaxies is relatively small (~10%) both in the Virgo cluster and the field, suggesting that starvation is not a major factor in the reduced star formation rates of Virgo spiral galaxies. The majority of Virgo spiral galaxies have their Hα disks truncated, whereas truncated Hα disks are rarer in isolated galaxies (➋ *Fig. 5-6*). Most of the Hα-truncated galaxies have relatively undisturbed stellar disks and normal to slightly enhanced inner disk star formation rates. Koopmann and Kenney (2004a, b) suggested that ram pressure stripping is the main mechanism causing the reduced star formation rates of Virgo spiral galaxies. Fumagalli and Gavazzi (2008) found in their much smaller Virgo cluster galaxy sample a larger fraction of truncated/anemic galaxies.

7 The Global Picture

Based on the observational findings presented in ➋ Sect. 5, a consistent global picture for galaxy evolution in groups and cluster can be constructed.

Since in galaxy groups, the galaxy velocity dispersion (~200 km s^{-1}) and the density of the intergalactic medium are low, ram pressure is in most cases negligible and slow galaxy–galaxy collisions should be the main driver of galaxy evolution. Close flybys leading to important tidal perturbations are relatively frequent. Tidal interactions can pull gas and stars out of the gravitational potential of a galaxy. In X-ray bright groups, which have a rich intergalactic medium, the tidally expelled gas might then be stripped by ram pressure. The merging of two spiral galaxies leads to a lenticular or elliptical galaxy, depending on the interaction parameters. The end product are galaxies with a quenched star formation which then evolve passively. Thus, group galaxies can already be gas deficient with a low star formation rate with respect to their field counterparts and/or of early type before they fall into a galaxy cluster. This phenomenon is termed as preprocessing of galaxies in groups. Cortése et al. (2006) presented a nice illustration of a compact group with a bright lenticular galaxy which is falling into the galaxy cluster Abell 1367. If the low-luminosity group galaxies were observed later, when ram pressure has stripped their gas entirely and star formation has stopped, they would probably resemble the post-starburst galaxies detected by Poggianti et al. (2004) in the Coma cluster.

Once a galaxy enters the cluster, its fate depends on the eccentricity of its orbit and the distribution of the intracluster medium. Observations indicate that HI-deficient galaxies are on more eccentric orbits within a cluster than HI-normal galaxies (Dressler 1986; Solanes et al. 2001). Numerical simulations (Ghigna et al. 1998) showed that galaxy dark matter halos evolving within a cluster settle on isotropic orbits with a median ratio of pericentric to apocentric radii of 1:6. In dynamically young, spiral-rich clusters as Virgo, the intracluster medium distribution is peaked on the central elliptical galaxy. In relaxed, spiral-poor clusters such as Coma, the intracluster medium distribution is more extended. In the latter case, less eccentric orbits are



■ **Fig. 5-6**

Left: **R-band and Hα images of three Virgo cluster galaxies with severely truncated star-forming disks. These galaxies have regular stellar disks, and their Hα morphologies show symmetric rings of star formation near the truncation radius. *Middle*: Hα and R-band radial profiles. *Right*: cluster radial distributions of each galaxy class, plotted as percentages of all the galaxies in that radial bin. Bins are ≤3°, 3° – 4.5°, and ≥4.5°. There is a clear radial dependence for the normal and truncated/normal classes, with fewer normal and more truncated/normal galaxies closer to the cluster center. Other Hα classes have a flatter distribution (but contain fewer sample galaxies; from Koopmann and Kenney (2004b))**

needed to remove a significant amount of gas from the galaxy by ram pressure stripping. The velocity dispersion of cluster galaxies is typically higher than 500 km s^{-1}. In the cluster core, where the galaxy density is high, gravitational encounters occur in the form of rapid flybys. These lead only to significant distortions if the impact parameter is small. At the apocenter of a galaxy orbit, far away from the cluster center, a slow galaxy–galaxy encounter is not excluded. The primary effect of the cluster environment on a galaxy is the quenching of its star formation. If ram pressure is the cause of the quenching, the timescale is small (10–100 Myr). If the main cause for the halt of star formation are gravitational interactions, the timescale is much longer ≥1 Gyr. The high fraction of spiral galaxies with truncated gas and star formation disks indicates that ram pressure is an important environmental effect in the Virgo cluster. Once the galaxies have lost their outer gas disk, star formation stops in these regions. There is no direct observational evidence that ram pressure instantaneously alters the gas content of the inner disk in a significant way. The molecular gas contents and star formation rates of stripped galaxies are similar to those of unperturbed galaxies. After a few Gyr, the gas of the inner disk will be consumed by star formation and truncated gas, and star formation disk will become anemic.

In the absence of significant star formation, the galaxy's color becomes red and the luminosity-weighted mean stellar population age increases. For the galaxy population of the local universe, a change in the fraction between the number of blue young and red old galaxies can already be observed in regions of low galaxy density. As expected, this fraction decreases with increasing density. In cluster of galaxies, the fraction between the number of blue young and red old galaxies is reversed with respect to that of isolated galaxies. Old red galaxies are the dominant population in clusters (➋ *Fig. 5-7*; Blanton and Moustakas 2009). At fixed morphological class according to light concentration, the distribution of stellar population ages strongly depends on environment. In contrast, galaxy–scaling relationships keep constant with environment, i.e., the distribution of galaxy morphologies does not depend strongly on environment once color is fixed (e.g., Bamford et al. 2008; Blanton and Moustakas 2009). Based on the analysis of large samples of nearby galaxies, it is claimed that environmental effects are relatively local. It appears that preprocessing of galaxies in groups is the main driver of these effects (Blanton and Moustakas 2009). However, this does not exclude that cluster environment can significantly affect infalling galaxies (see ➋ Sect. 8.2).

How can a spiral be transformed into a lenticular galaxy? Dressler (1980) showed that the bulge sizes and bulge-to-disk ratios of lenticular galaxies are systematically larger than those of spiral galaxies. Boselli and Gavazzi (2006) confirmed this result for the Virgo cluster. Christlein and Zabludoff (2004) excluded the generation of early-type galaxies (E/S0) by fading the disks of late-type galaxies. Under their assumptions, the bulge luminosities needed to be physically enhanced for such a transformation. Since ram pressure does not act on the stellar content of a spiral galaxy, only tidal interactions can lead to such an enhancement. Moreover, (i) the large scatter in the Tully–Fisher relation of Coma and Virgo cluster S0 galaxies in comparison to the known late-type spiral relation, together with the small zero-point offset, and (ii) the weak dependence of morphological segregation on galaxy density cannot be explained by an S0 formation through simple gas removal from spiral galaxies. Again, tidal interactions like slow encounters, minor mergers, or harassment are needed to transform spiral galaxies into lenticulars (Dressler and Sandage 1983; Neistein et al. 1999; Hinz et al. 2003).

Bright and faint lenticulars might not share the same history (Barway et al. 2009). Bright cluster and field lenticulars resemble ellipticals and bulges of early-type spirals suggesting that they may have formed at early epochs via major mergers or rapid collapse. Faint cluster lenticulars show systematic differences with respect to faint field lenticulars. These differences support

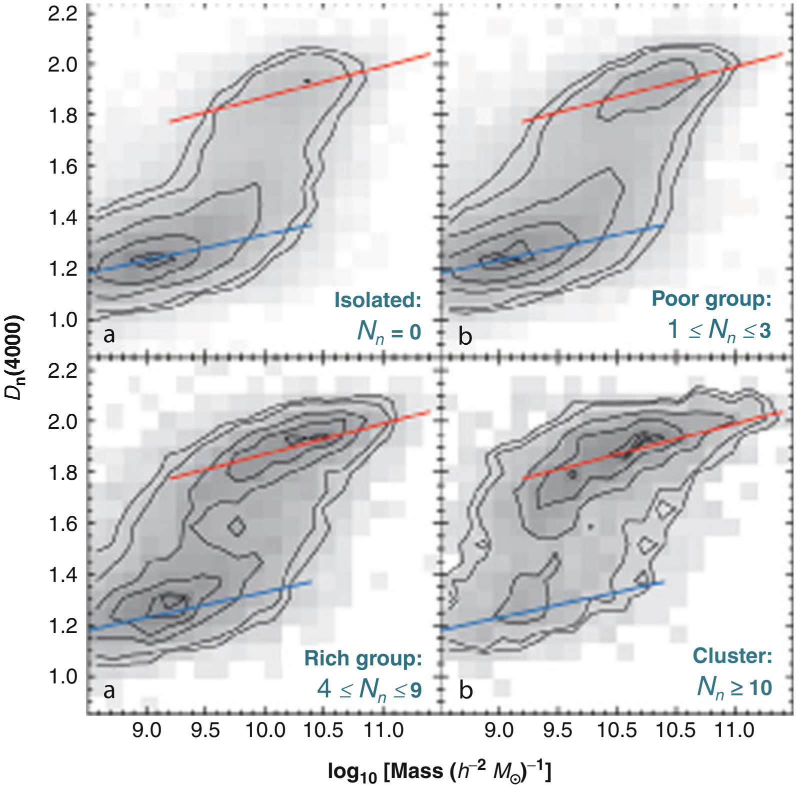


Fig. 5-7

Distribution of the stellar age measured by the 4, 000 Å break (defined by Balogh et al. 1999) as a function of the stellar mass for different environments (defined according to the number of neighboring galaxies with $M_r - 5 \log_{10} h < -18.5$ within a projected radius of 500 h^{-1} kpc and a velocity of 600 km s^{-1}). The *blue* and *red* lines are references for the young and old galaxy populations (From Blanton and Moustakas 2009)

the idea that the bulge and disk components fade after the galaxy falls into a cluster, while simultaneously undergoing a transformation from spiral to lenticular morphologies (Barway et al. 2009; Boselli and Gavazzi 2006). It thus seems that, if one wants to look for the end product of spiral galaxy transformation in a galaxy cluster, one should have a look at faint cluster lenticulars.

Due to their shallow gravitational potential, dwarf galaxies are particularly vulnerable to environmental interactions. Gravitational interactions, most probably in form of galaxy harassment (➋ Sect. 3.1), can efficiently remove stars and gas from the outer parts of the galaxy and transform the galaxy's morphology. Ram pressure stripping (➋ Sect. 3.2) of dwarf irregular galaxies (dI) removes the gas, stops star formation, and makes the galaxy evolve passively.

The density morphology relation (➋ Sect. 2) extends to low-luminosity dwarf galaxies (Binggeli et al. 1988, 1990). *Early-type dwarf elliptical* (dE) galaxies strongly prefer dense environments. They represent the numerically dominant galaxy population in nearby galaxy clusters and groups (Ferguson and Binggeli 1994). In galaxy groups, dEs are preferentially found near giant galaxies. One of the specific properties making dE/dS0s different from late-type dwarf irregular and spiral galaxies is the lack of interstellar medium (ISM) and, hence, ongoing star formation. Stellar populations of dE galaxies are remarkably different from those of giant early-type galaxies pointing out to differences in star formation histories and chemical evolution.

Most dE galaxies have compact nuclei with a whole range of sizes and central surface brightnesses (Côté et al. 2006). Based on their analysis of 413 Virgo cluster dwarf ellipticals, Lisker et al. (2007) divided their sample into nucleated and nonnucleated dEs depending on the presence of a strong nucleus compared to the galaxy's total luminosity. The nonnucleated dEs can show disk features like bars or spiral arms (also named dwarf S0 galaxies, dS0) and can have central star formation. Whereas the nucleated dEs have a centrally peaked distribution within the cluster-like giant elliptical and lenticular galaxies, the distribution of nonnucleated dEs shows no central clustering.

About half of the dEs are supported by rotation (van Zee et al. 2004a; Toloba et al. 2011). Based on the observed maximum rotation velocities, the rotating dwarf galaxies appear to follow the Tully–Fisher relation for gas-rich dwarf and spiral galaxies. No significant difference in dominant stellar populations between rotating and nonrotating dwarf elliptical galaxies (Geha et al. 2003; van Zee et al. 2004b) or between dEs with and without disks (Paudel et al. 2010) were found. The analysis of the color–magnitude relation of these objects led Lisker et al. (2008) to the conclusion that there must be multiple formation channels. Boselli et al. (2008) compared UV to radio centimetric properties of star-forming and quiescent Virgo dwarf galaxies to the predictions of multizone chemospectrophotometric models. The models include the quenching of star formation due to ram pressure stripping. These authors suggested that young, low-luminosity, high surface brightness star-forming galaxies such as late-type spirals are probably the progenitors of relatively massive dwarf ellipticals, while it is likely that low surface brightness Magellanic irregulars evolve into very low surface brightness quiescent objects hardly detectable in ground-based imaging surveys.

Dwarf elliptical thus represent a heterogeneous class of galaxies. Most probably several different mechanisms, including environmental effects as galaxy harassment and ram pressure stripping, are involved in the creation of the overall population of dEs.

Ultracompact dwarf galaxies (UCDs) have properties between dwarf ellipticals and globular clusters. They were first discovered in the Fornax and Virgo clusters (Hilker et al. 1999; Drinkwater et al. 2000; Haşegan et al. 2005). They are also found around giant galaxies in nearby groups. Ultracompact dwarfs are characterized by predominantly old stellar populations (≥8 Gy, e.g., Chilingarian et al. 2011), small sizes (half-light radii of $10 \le r_h \le 100\,\mathrm{pc}$), and dynamical masses of $2 \times 10^6 \le M \le 10^8\,\mathrm{M}_\odot$ (Hilker et al. 2007; Mieske et al. 2008). As dEs, ultracompact dwarfs represent a heterogeneous class of objects (Mieske et al. 2006). They might be (i) very massive globular clusters, (ii) tidally stripped nucleated dEs (Bekki et al. 2001), or (iii) end products of small-scale primordial density fluctuations in dense environments (Phillipps et al. 2001). It is suggested that the most massive ultracompact dwarfs are remnants of more extended galaxies, whereas the less massive ones represent a transition objects toward the regime of ordinary globular clusters (Chilingarian et al. 2011).

If tidal stripping acts on massive progenitors, one would expect the remnants to be larger and more luminous than UCDs. An example of such an object is M 32, a *compact elliptical* (cE) satellite of the Andromeda Galaxy. It has a luminosity comparable to typical dE in clusters but

10 times smaller half-light radius, therefore 1,000 times higher stellar density per unit of volume. For several decades cE galaxies were considered unique since only three of them including M 32 were known and several dedicated searches failed to detect any. Therefore, even solid arguments for M 32 to be a heavily tidally stripped lenticular galaxy (Graham 2002) did not allow to consider tidal stripping an important channel of galaxy evolution because of the lack of examples. The situation might change with the discovery of a population of 21 cE galaxies in nearby galaxy clusters by Chilingarian et al. (2009). These authors showed that the properties of the cE galaxies can be reproduced by numerical simulations of tidal stripping.

8 Resolved Multiwavelength Interaction Diagnostics

Can we catch a galaxy with an ongoing environmental interaction? To do so, imaging observations with a spatial resolution of ~1 kpc are necessary. With multiwavelength imaging observations, the reaction of the multiphase interstellar medium and star formation to these interactions can be studied in detail. Deep optical imaging reveals perturbations of the galaxy's stellar content due to tidal interactions. The comparison between optical and interferometric HI observations can discriminate between tidal and hydrodynamic interactions. Since ram pressure only affects the gas, a symmetric stellar disk with a truncated gas disk within the optical radius and a one-sided gas tail in a cluster galaxy are signs of ongoing ram pressure stripping (e.g., Chung et al. 2007). In rare cases, extraplanar molecular gas traced by its CO emission can also be found (Combes et al. 1988; Vollmer et al. 2005b, 2006). Perturbed and extraplanar star formation distributions are observed in Hα (e.g., Koopmann and Kenney 2004b) or UV (e.g., Abramson et al. 2011). Sometimes the radio continuum halo can be compressed on the side where ram pressure is acting (Gavazzi et al. 1995; Crowl et al. 2005). In rare cases, an extended one-sided X-ray tail is detected (e.g., Sun et al. 2006).

Relatively new diagnostic tools are the polarized radio continuum and radio/FIR distributions of cluster spiral galaxies.

Under the assumption of a constant FIR-radio correlation within normal galaxies, Murphy et al. (2009) used Spitzer infrared data to create model radio maps, which were compared to observed radio images. These authors found that galaxies, which are affected by ram pressure stripping, have enhanced global radio fluxes with respect to the FIR. These galaxies contain regions along their outer edges where the observed radio surface brightness is significantly below the model expectation (➋ *Fig. 5-8*). The radio-deficient regions are located in the direction of the ram pressure wind and often show an enhanced polarized radio continuum emission (see ➋ *Fig. 5-9* for NGC 4522).

Polarized radio continuum emission is due to relativistic electrons with density n_e gyrating around the regularly oriented, large-scale magnetic field B: $S_\mathrm{PI} \propto n_\mathrm{e} B^2$. The polarized radio continuum emission is enhanced in regions where shear and compression of the regular or random magnetic field component parallel to the sky plane occur. From spectroscopic observations, noncircular motions of the order of ~10 km s^{-1} induced by an interaction can be determined by a detailed analysis of a galaxy's velocity field (e.g., Schoenmakers et al. 1997). On the other hand, the distribution of polarized radio continuum emission represents a very sensitive tool for uncovering the transverse motions of the ISM (Beck 2005) even in the case of unfavorable inclinations (close to face-on). Therefore, the information contained in polarized radio continuum emission is complementary to that of Hα, CO, and HI observations. The total

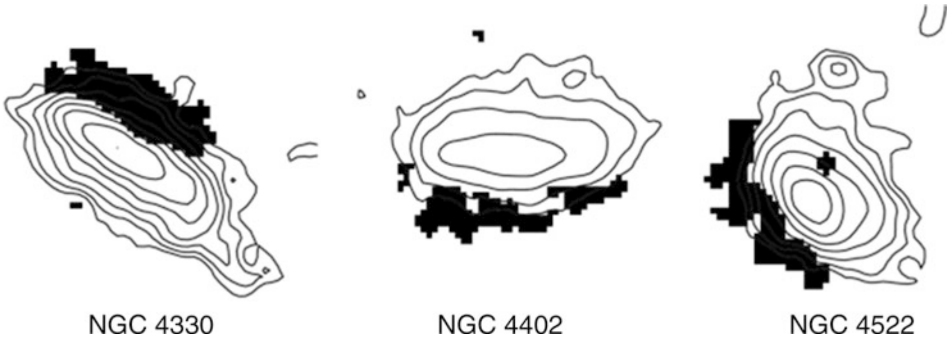


Fig. 5-8
Radio-deficient regions of the Virgo cluster spiral galaxies NGC 4330, NGC 4402, and NGC 4522 with radio continuum contours (From Murphy et al. 2009)

radio continuum emission is sensitive to the turbulent small-scale magnetic field, which is usually a factor of 2–5 larger than the regular large-scale magnetic field in spiral arms and 1–2 times larger in the interarm regions at a typical resolution of a few 100 pc (Beck 2001). Whenever there is enhanced turbulence due to an enhanced star formation efficiency, the large-scale magnetic field is diminished. The polarized radio continuum emission has to be observed at a frequency that is high enough to avoid significant Faraday rotation (typically >2 GHz). In a pioneering work, Vollmer et al. (2007) presented 6 cm polarized radio continuum emission of 8 Virgo spiral galaxies. All galaxies show strongly asymmetric distributions of polarized intensity with elongated ridges located in the outer galactic disk. These features are not found in existing observations of polarized radio continuum emission of field spiral galaxies, where the distribution of 6 cm polarized intensity is generally relatively symmetric and strongest in the interarm regions. Once an asymmetric distribution of polarized radio continuum emission is detected, one has to discriminate between shear or compression motions as its cause. In the case of ram pressure compression, one expects an HI gas tail at the opposite side of the compression region (e.g., Vollmer et al. 2004b; *Fig. 5-9*). Tidal interactions can lead to compression and shear motions (Soida et al. 2006).

8.1 Environmental Effects in Nearby Galaxy Groups

In galaxy groups, encounters between two galaxies are the most important environmental effect. Relative velocities between galaxies are small, and for small impact parameters, tidal fields produce important distortions of the stellar and gaseous content of the galaxies. Since galaxy density is highest in compact groups, galaxy encounters occur frequently in this environment. One of the best studied compact group is Stephan's Quintet which is made of four group galaxies and one foreground object. A fifth galaxy, situated ~4′ to the east, is also dynamically associated to the group. The group has experienced a violent dynamical history with numerous interactions between the different members during the past Gyr. As a result of these interactions, two tidal arms, a faint older one and a brighter young one stemming from NGC 7319, have been created toward the eastern side of the group (Sulentic et al. 2001). The other group spiral galaxies show important tidal perturbations. Because of these interactions, most of the gas is found

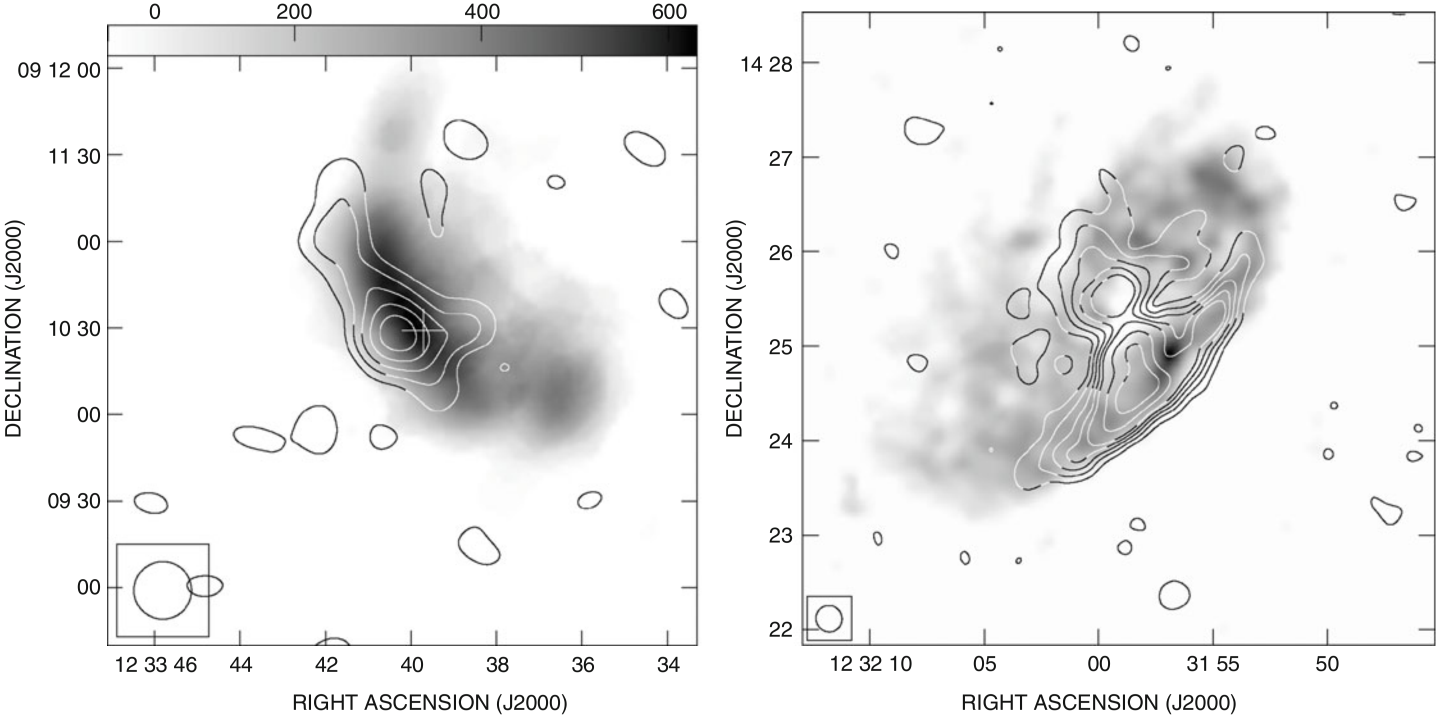


Fig. 5-9

6-cm polarized radio continuum emission (*contours*) on the Hɪ distribution (*greyscale*) of the ram pressure stripped Virgo spiral galaxies NGC 4522 (*left*; from Vollmer et al. 2004b) and NGC 4501 (*right*; from Vollmer et al. 2010). The regions of gas compression visible in the polarized radio continuum are on the opposite side of the extraplanar Hɪ or Hɪ extensions

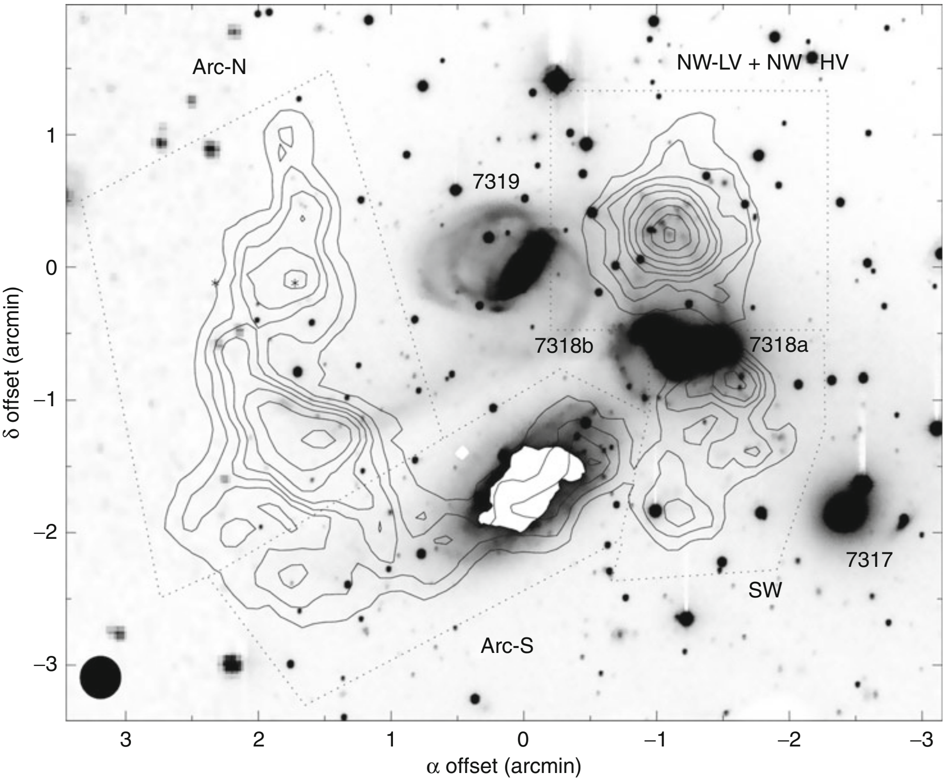


Fig. 5-10
Stephan's Quintet. Map of the total HI column density distribution (contours) superposed on the R-band image (from Williams et al. 2002). The galaxy with the white area (NGC 7320) is in the foreground. The HI projected onto this galaxy actually belongs to the group

in a highly disturbed intragroup medium (*Fig. 5-10*; Williams et al. 2002). The last and presumably ongoing event involves the collision of the gas-rich spiral NGC 7318b with the debris field produced by past interactions.

Due to the dense environment, galaxy and group evolution is accelerated in compact groups. In Stephan's Quintet, we can directly observe galaxy transformation via multiple gravitational encounters. In this way, gas and stars are loosened and eventually removed from the parent galactic disks. The fraction of the diffuse light from tidally stripped stars to the total light from the group can be up to ~50% in compact groups (da Rocha and Mendes de Oliveira 2005). Galaxy mergers lead to lenticular or elliptical galaxies. Indeed, there are compact groups whose galaxy populations are dominated by early-type galaxies (e.g., Huchtmeier 1997). Since gas disks of normal spiral galaxies are more extended than stellar disks, the amount of expelled gas is larger than that of stars. This gas then forms an intragroup medium. It is expected to be heated through evaporation in a previously existing intragroup medium or through large-scale shocks from freshly accreting material falling into the galaxy group.

To investigate the importance of ram pressure stripping in compact groups, Rasmussen et al. (2008) analyzed the diffuse X-ray emission of 8 HI-deficient Hickson compact groups.

If ram pressure stripping were the dominant cause of HI removal, one would expect an important intragroup medium traced by X-ray emission in these groups. Their finding that the most HI-deficient groups do not show a detectable hot intragroup medium suggests that tidal interactions are the most likely cause of gas removal. This is consistent with the absence of a correlation between the fraction of HI-deficient galaxies in a group and the X-ray luminosity in the sample of Kilborn et al. (2009). On the other hand, HI imaging of spiral galaxies in X-ray bright groups (Sengupta et al. 2007) revealed disturbed morphologies and truncated gas disks compared to field spirals. The observed gas disk truncation might be an indication that ram pressure stripping plays a role in groups with a prominent intragroup medium.

Thus, for the overall statistics tidal interaction dominate and ram pressure seems to play a minor role in the evolution of group spiral galaxies. However, in X-ray bright groups, it may represent an additional cause for the observed HI deficiency. Most probably, the ISM has to be loosened tidally to be stripped efficiently by ram pressure. In rare cases, arriving spiral galaxies can undergo important ram pressure if the intragroup gas is still in the form of relatively dense tidal tails of atomic gas from a previous galaxy–galaxy encounter. As an example, Clemens et al. (2000) presented the case of ram-pressure stripping by a gaseous tidal tail in the interacting pair of galaxies NGC 4490/4485.

8.2 Environmental Effects in Nearby Galaxy Clusters

Multiwavelength imaging observations of cluster galaxies give insights when and where in the cluster a galaxy experiences an environmental interaction and how its different components are affected by this interaction. In the following, we will have a look at five close galaxy clusters: Coma, Norma, Abell 1367, Virgo, and Fornax. The five clusters represent quite different environments. The Coma, Norma, and Abell 1367 clusters are more massive than the Virgo and Fornax clusters. Whereas the Norma and Virgo clusters are dynamically active with galaxy subgroups or clusters falling onto and merging with the main cluster, Coma and Fornax are more dynamically quiescent, i.e., their galaxy accretion rates are much lower than those of Norma, Abell 1367, and Virgo. However, even in the more quiet systems galaxies and galaxy groups are still falling into the main cluster. The fraction of HI-deficient galaxies is large in Coma, Virgo, and Norma and intermediate in Abell 1367. Galaxies in the Fornax cluster do not show signs of strong HI deficiency (Horellou et al. 1995). In the following, the properties of these clusters sorted by velocity dispersion are presented.

1. The *Coma cluster* is a massive relaxed cluster with a mass of ~$5 \times 10^{14}\,M_\odot$ within 1 Mpc and galaxy velocity dispersion of 950 km s^{-1}. It is the richest nearby (~90 Mpc) cluster and represents the end product of an ancient merger of two clusters. The dominating two central elliptical galaxies, NGC 4874 and NGC 4889, are reminiscent of this merger. The Coma cluster is spiral-poor which is typical for rich clusters. Despite the symmetric galaxy distribution, a galaxy group in the southwest of Coma, associated with the giant elliptical galaxy NGC 4839, has probably passed the cluster core in the recent past. The cluster is X-ray luminous, has a symmetric main component with an extended intracluster medium distribution, and has some substructure related to galaxy groups (Briel et al. 1992). For a review on the Coma cluster, see Biviano (1998).
2. The *Norma cluster* (Abell 3627) is the nearest (~65 Mpc) rich, massive cluster of galaxies (Kraan-Korteweg et al. 1996; Woudt 1998), with properties comparable to the Coma cluster

(Mazure et al. 1998). It has remained relatively unexplored in comparison to its well-known counterpart, mainly because of its location at low galactic latitudes in the southern zone of avoidance (ZOA) where dust extinction and star crowding dilute its appearance. The dynamical mass of the Norma cluster is ~10^{15} $M_\odot$ within its Abell radius of 2 Mpc. The galaxy velocity dispersion is 925 km s^{-1} (Woudt et al. 2008). The spiral/irregular galaxies reveal a large amount of substructure. The X-ray emission distribution is not spherically symmetric and shows indications of an ongoing cluster merger (Böhringer et al. 1996).

3. *Abell 1367* forms together with the Coma cluster the Coma-Abell 1367 supercluster at a distance of ~90 Mpc. Its velocity dispersion is 880 km s^{-1} (Moss et al. 1998). The cluster is a rather unusual example of a rich cluster with a comparatively high fraction of spiral galaxies. It has been identified as having significant optical and X-ray substructure (Grebenev et al. 1995). The X-ray emission is elongated along a southeast-northwest axis and contains small, localized clumps. This is interpreted as a merger of two subclusters (Donnelly et al. 1998). The high spiral fraction and a relatively cool intracluster medium temperature (Donnelly et al. 1998) are typical of what is expected for a dynamically young system.
4. The *Virgo cluster* is less massive (~2 × 10^{14} $M_\odot$ within 1.5 Mpc) and has a galaxy velocity dispersion of ~600 km s^{-1}. It represents the nearest cluster (17 Mpc) in the northern hemisphere. The Virgo cluster is spiral-rich and has a lot of substructure (see, e.g., Schindler et al. 1999). Different galaxy groups can still be distinguished spatially and kinematically from the main cluster. Virgo is thus said to be dynamically young. Its X-ray luminosity is ~6 times smaller than that of the Coma cluster. X-ray emission can be associated to the main cluster, where it is strongly peaked on the central elliptical galaxy M 87, and several substructures (Böhringer et al. 1994).
5. After Virgo, the largest concentration of galaxies within 20 Mpc is the *Fornax cluster*. Its spiral-rich galaxy distribution has a more regular shape than that of the Virgo cluster, indicating a more dynamically evolved state. Fornax is also considerably smaller and denser than Virgo, with a core radius ~40% that of Virgo and a central density twice as large. The total mass of Fornax is 0.7–2 × 10^{14} $M_\odot$ within ~1 Mpc (Drinkwater et al. 2001; Dunn and Jerjen 2006), its velocity dispersion 380 km s^{-1} (Drinkwater et al. 2001). The X-ray emission shows an asymmetric spatial distribution, and the central elliptical galaxy, NGC 1399, is offset from the center (Paolillo et al. 2002), which may be related to large-scale dynamical evolution such as infall motions of galaxies into the cluster (Dunn and Jerjen 2006).

In a cosmological context, galaxies assemble into small groups which then merge forming large groups or small clusters like Fornax. A central cluster is growing with the accretion of groups and small clusters (Virgo). When two big clusters merge, they have the appearance of Abell 1367 or the Norma cluster. Without further accretion, the cluster virializes and becomes spherical (Coma). Galaxy clusters can be classified according to their degree of relaxation. Non-relaxed clusters have an irregular overall shape, are spiral-rich, and have a low X-ray luminosity. On the other hand, relaxed clusters are spherical, rich of early-type galaxies, and have a high X-ray luminosity. However, not all clusters follow these rules. Each cluster is different and has its own personality according to its assembling history. This cosmological scenario is further discussed in ➲ Sect. 10.

In the next section, we will have a close look at resolved multiwavelength observations of cluster galaxies. Early-type cluster galaxies have a corona of abundant hot gas (several 10^6 K), which is in hydrostatic equilibrium with the galaxy's gravitational potential and emits in X-rays. In addition, small amount of cold gas ~10^8 $M_\odot$ are often found in the galaxy cores. The ISM of late-type galaxies has a multiphase structure with a dominant neutral warm and cold phase as described in ➤ Sect. 5. It is confined within the galactic disk and supported by rotation. Dwarf spiral and irregular galaxies possess a neutral warm/cold ISM which is either supported by rotation and velocity dispersion, respectively. Due to the shallow potential well of these galaxies, their gas content can be removed easily by environmental interactions. Because of these differences, early-type, late-type, and dwarf galaxies will be discussed separately.

8.2.1 Early Type Galaxies

In early-type cluster galaxies, the hot corona is exposed to the intracluster medium. Environmental effects are thus detectable in X-ray emission. The signpost of ram pressure compression is a sharp edge on one side of the galaxy. The X-ray surface brightness drops by an order of magnitude over the edge. These "cold fronts" are contact discontinuities, in which a sharp drop in surface brightness (gas density) is accompanied by a corresponding rise in gas temperature. This is characteristic for the motion of a spheroid through a uniform gas and is consistent with a jumplike density discontinuity at the boundary in the direction of the galaxy's motion within the intracluster medium. The sharpness of the edge is a strong function of the inclination angle between the galaxy's 3D velocity vector and the plane of the sky. The edge is most prominent if the galaxy has a negligible radial velocity with respect to the surrounding intracluster medium. Sometimes, an X-ray low surface brightness tail can be found at the opposite side of the sharp edge (e.g., Machacek et al. 2006).

X-ray spectra yield information on the density, temperature, and metallicity of the interstellar medium of the elliptical galaxy and the intracluster medium. The ISM density and temperature are typically $n_{\mathrm{ISM}} \sim$ a few $10^{-3}\,\mathrm{cm}^{-3}$ and $T \sim 0.5\,\mathrm{keV} = 6 \times 10^{6}\,\mathrm{K}$ with an about solar metallicity. The density and metallicity of the intracluster medium are smaller, and its temperature is higher ($T_{\mathrm{ICM}} \geq 1.5\,\mathrm{keV} \sim 2 \times 10^{7}\,\mathrm{K}$). The pressure jump can be used to determine the Mach number M_1 and thus the velocity with which the galaxy moves through the intracluster medium (Landau and Lifshitz 1959). For a subsonic motion ($M_1 \leq 1$):

$$\frac{p_0}{p_1} = \left(1 + \frac{(\gamma - 1)}{2} M_1^2\right)^{(\gamma/\gamma-1)}, \tag{5.9}$$

for supersonic motions ($M_1 > 1$):

$$\frac{p_0}{p_1} = \left(\frac{\gamma + 1}{2}\right)^{(\gamma+1)/(\gamma-1)} M_1^2 \left(\gamma - \frac{\gamma - 1}{2M_1^2}\right)^{-1/(\gamma-1)}, \tag{5.10}$$

where $\gamma = 5/3$ is the adiabatic index for a monoatomic ideal gas and p_0/p_1 the gas pressure in the ISM and the intracluster medium, respectively. Together with the observed radial velocity of the galaxy, the transverse velocity in the plane of the sky and the inclination angle between the galaxy's 3D velocity and the plane of the sky can be determined. The latter has to be consistent with the observed X-ray morphology.

With this method, Machacek et al. (2005) derived a pressure jump of $p_0/p_1 \sim 2$ implying a Mach number of $M_1 \sim 1$ for the elliptical galaxy NGC 1404 in the Fornax cluster (➤ *Fig. 5-11*). This galaxy is approaching the dominant elliptical galaxy of the Fornax cluster, NGC 1399, and undergoes significant ram pressure by the Fornax cluster intracluster medium.

Another example of a cluster elliptical galaxy undergoing active ram pressure is NGC 4552 (M 89) in the Virgo cluster. This galaxy is located at a projected distance of $72' = 360\,\mathrm{kpc}$ east

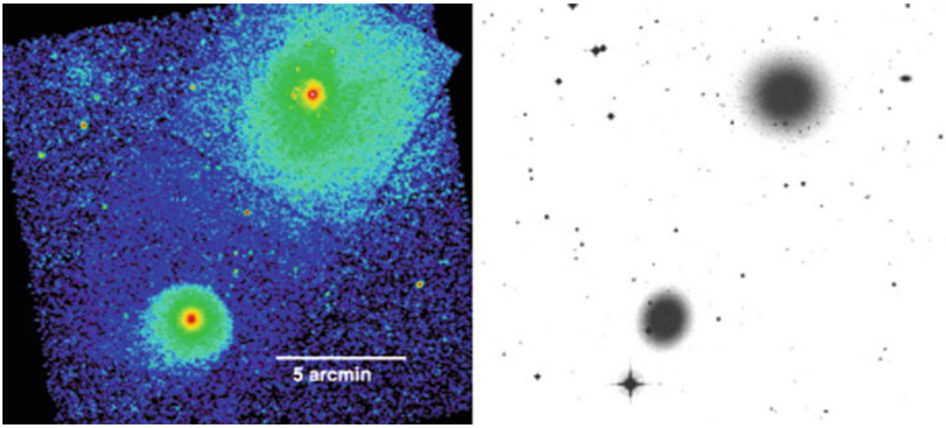


Fig. 5-11
Chandra ACIS X-ray (*left*) and DSS (*right*) images of NGC 1404 (southeast, *lower left*) and NGC 1399 (northwest, *upper right*) galaxies in the Fornax cluster (From Machacek et al. 2005)

of M 87, the central elliptical galaxy of the Virgo cluster. NGC 4552 has an extended (~10 kpc) X-ray low surface brightness tail (Machacek et al. 2006). The properties of the X-ray tail are consistent with it being composed primarily of ram pressure stripped galaxy gas. The tail is denser and cooler than the Virgo intracluster medium. On the opposite side, a classical cold front is detected. Galaxy gas inside the leading edge is cool ($kT = 0.43$ keV) compared to the surrounding 2.2-keV Virgo intracluster medium. The pressure ratio of ~7 across the leading edge of the ram pressure interaction corresponds to a Mach number of ~2. Two horns of emission extending 3–4 kpc to either side of the edge are composed of gas in the process of being stripped from the galaxy due to the onset of Kelvin-Helmholtz instabilities. This galaxy thus also undergoes turbulent viscous stripping (see Sect. 3.2).

The same overall X-ray morphology is found in another massive Virgo elliptical galaxy, NGC 4472 (M 49; Biller et al. 2004), which is located at a projected distance of 1.3 Mpc south of M 87. A taillike structure in the X-ray emitting gas extends $\sim 8' = 36$ kpc to the southwest of the galaxy core. The northeastern edge in the X-ray emission distribution is most probably the result of ram pressure compression as NGC 4472 falls toward the center of the Virgo cluster.

The most enigmatic elliptical galaxy in the Virgo cluster is M 86. The second brightest Virgo elliptical galaxy is located at a projected distance of only 350 kpc from M 87 and has a high radial velocity with respect to the Virgo cluster mean (~1,000 km s^{-1}). It might be located as far as 1 Mpc behind M 87 (Mei et al. 2007). An X-ray study of M 86 with Einstein (Forman et al. 1979) revealed a peak of emission centered on M 86 and a plume extending northwest of the galaxy. Subsequent observations showed substructure in the M 86's X-ray halo. Finoguenov et al. (2004a) revealed a cold front in the southwest of the galaxy core which is most probably due to ram pressure compression. They ascribed the huge X-ray plume to a gravitational interaction. This scenario has gained support by the detection of faint Hα emission connecting the perturbed spiral galaxy NGC 4438 to M 86 (*Fig. 5-17*; Kenney et al. 2008). On the other hand, Randall et al. (2008) suggested that the X-ray plume is due to ram pressure stripping.

8.2.2 Late Type Galaxies

Tidal interactions lead to asymmetries in the stellar distribution of the galactic disk observable in the NIR, optical, and UV. The signposts of ram pressure stripping are (i) gas disks which are truncated inside the optical radius, (ii) one-sided HI tails together with a symmetric old stellar disk, and (iii) radio-deficient regions and/or asymmetric ridges of polarized radio continuum emission located in the outer-disk opposite to the gas tail (➋ *Fig. 5-9*). One-sided tails of warm or hot gas observable in Hα and X-rays are much rarer than tails of atomic gas. Radio continuum tails are the exception. A significant fraction of perturbed cluster galaxies undergo a tidal interaction and ram pressure stripping at the same time. In the following, an inventory of environmentally affected disk galaxies in the different nearby clusters is presented.

Coma Cluster

Bravo-Alfaro et al. (2000) imaged the 19 brightest spiral galaxies in the *Coma cluster* in the HI line (➋ *Fig. 5-12*). The galaxies, which are located at projected distances smaller than $d = 0.6$ Mpc, show truncated gas disks with pronounced asymmetries in their atomic gas distributions and/or shifts between the optical and the HI positions. Twelve spiral galaxies were not detected with typical upper HI mass limits of $10^8\,M_\odot$. Seven of the twelve non-detections are located in the central region of Coma ($d < 0.6$ Mpc). In addition to HI asymmetries, UV and Hα asymmetries seem to be a common property of ram pressure stripped Coma cluster galaxies. Whereas the UV emission only traces young massive stars, the Hα emission can also provide from ionized dense gas. Yagi et al. (2010) found extended Hα clouds associated with 14 Coma cluster galaxies obtained from deep narrow-band imaging observations with the Subaru Telescope. The parent galaxies are blue and distributed farther than 0.2 Mpc from the peak of the X-ray emission of the cluster. Smith et al. (2010) found tails or trails of UV-bright debris in 13 star-forming Coma cluster galaxies, which they interpreted as young stars formed within gas stripped by ram pressure from the intracluster medium. Within 1 Mpc projected distance from the cluster center, about 30% of blue galaxies show UV trails. These trails are predominantly oriented away from the cluster center, indicating that the galaxies are falling into the cluster for the first time, along radial orbits.

One of the Coma spiral galaxies barely detected in HI is NGC 4848 (CGCG 160-055 on ➋ *Fig. 5-12*). This highly inclined spiral galaxy is located on the outermost X-ray contour northwest of the cluster center. The atomic gas is displaced to the northwest with respect to the optical disk. Its molecular gas distribution is also asymmetric with an off-center secondary maximum coincident with the inner part of the displaced atomic gas (Vollmer et al. 2001b). The Hα emission shows a double line in this conspicuous region. One line is due to galaxy rotation, the other line is most probably gas accelerated by ram pressure. Furthermore, an X-ray tail is detected to the northwest which is more extended than the HI emission (Finoguenov et al. 2004b). An estimation of the stripping radius (➋ 5.2) based on the intracluster medium density at the projected distance of NGC 4848 given by Briel et al. (1992) and a galaxy velocity of 2,000 km s^{-1} yields a stripping radius twice as large as the edge of the gas and star formation. Thus, either the galaxy already had its closest approach to the cluster center and is now leaving the cluster core as proposed by Vollmer et al. (2001a), or the intracluster medium has an overdensity or moves in the direction opposite to that of the galaxy's motion (see ➋ Sect. 9).

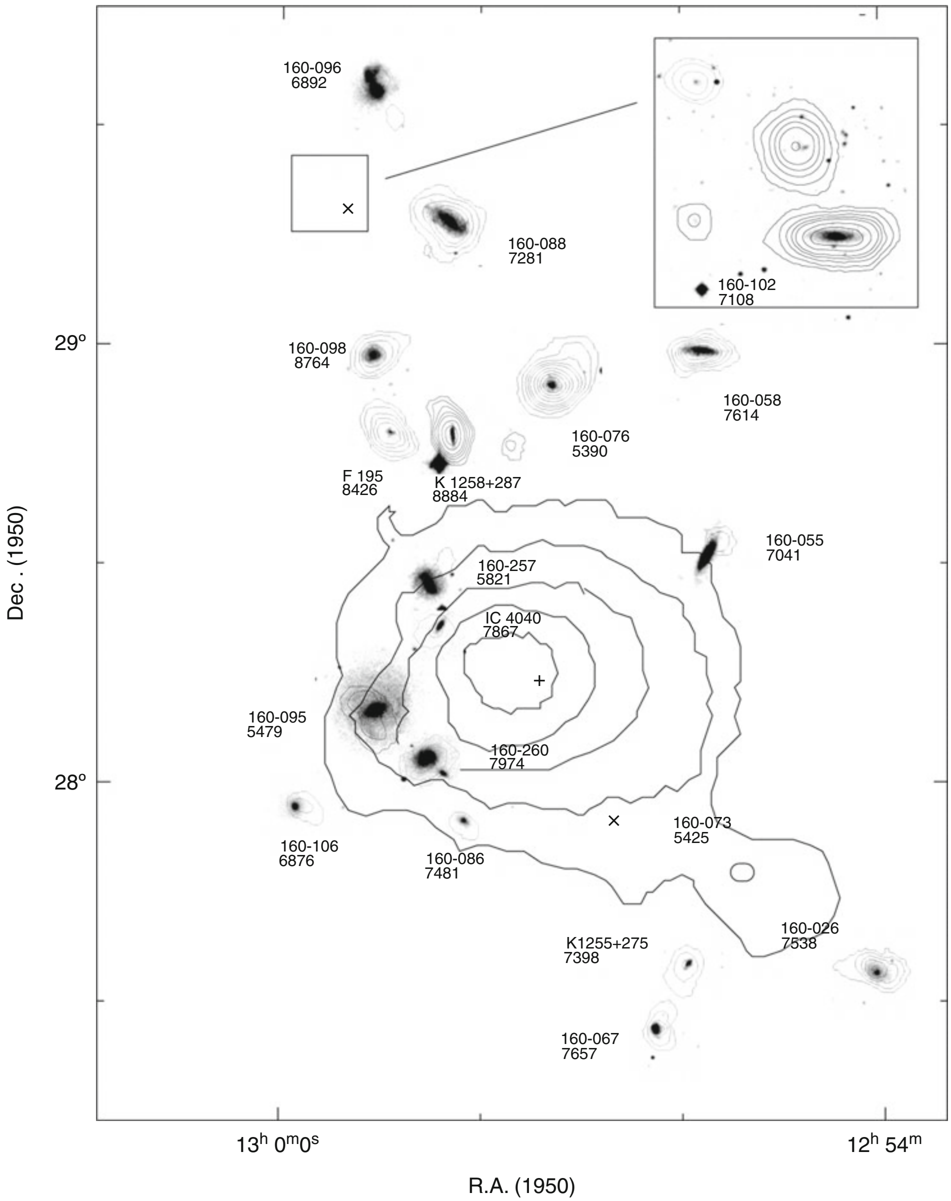


Fig. 5-12

Composite plot of individual HI maps of spiral galaxies of Coma observed with the VLA. Galaxies are shown at their proper position (except those in the rectangle, where the position of CGCG 160-102 is indicated with a times sign), and they are magnified by a factor of 7. The HI maps (*thin contours*) are overlaid on DSS optical images. The large-scale contours sketch the X-ray emission of the intracluster medium (From Bravo-Alfaro et al. 2000)

Norma Cluster

The detection of long one-sided X-ray tails is very rare compared to Hɪ detections. The most spectacular example is that of the late-type galaxy ESO 137-001 (Sun et al. 2006, 2007, 2010; ➋ *Fig. 5-13*), which is most probably stripped nearly face-on. The projected distance from the cluster's X-ray peak of this blue emission-line galaxy is only 180 kpc. About 80% of the galaxy's total X-ray emission corresponding to a gas mass of $\sim 10^9\ M_\odot$ is found behind the galactic halo. The X-ray tail consists of two components: a straight and thin (~8 kpc) main tail with a size of ~80 kpc and a secondary fainter and curved X-ray tail. The main tail is brightest in the middle and near the galactic disk. The gas temperature in the tail is constant ($kT \sim 0.8$ keV) and significantly lower than that of the surrounding intracluster medium ($kT \sim 6$ keV). The soft X-ray emission of the tail comes from the mixing of cold ISM with the hot intracluster medium. Diffuse Hα is associated with the X-ray tail. Close to the galactic disk (<20 kpc), Sivanandam et al. (2010) detected $2.5 \times 10^7\ M_\odot$ of shocked, warm molecular hydrogen. An important number of Hɪɪ regions are detected near or close to the X-ray tail up to a projected distance of ~40 kpc from the galactic disk. Thus, star formation proceeds in situ in the stripped material. The imprint of galactic rotation is still present in the velocity map of these Hɪɪ regions. It seems that the gas turbulence in the stripped gas tail did not greatly enhance the velocities of the extraplanar Hɪɪ regions.

A second 40-kpc-long X-ray tail has been detected behind another late-type galaxy, ESO 137-002 (Sun et al. 2010). Signature of gas stripping were also found in the Hα data, with a sharp Hα edge and an Hα tail extending to at least 20 kpc from the galactic nucleus. No Hɪɪ regions are found in this tail. The X-ray morphology points to a more edge-on stripping.

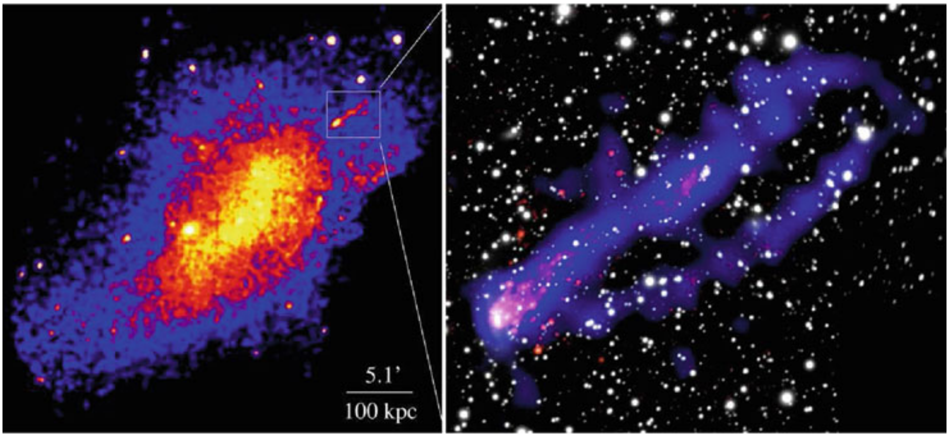


Fig. 5-13

***Left*: XMM-Newton 0.5–2 keV mosaic of the galaxy cluster Abell 3627. The main tail of ESO 137-001 is significant in the XMM-Newton image. *Right*: the composite X-ray/optical image of ESO 137-001's tail. The Chandra 0.6–2 keV image is in blue, while the net Hα emission is in red and the stellar image in white (From Sun et al. 2010)**

Abell 1367

The only disk galaxies with a long one-sided radio continuum tail are CGCG 97-073 and CGCG 97 079 in the *galaxy cluster Abell 1367* (Gavazzi et al. 1995). Faint Hα emission is associated with these tails. The sizes of the tails are 50–75 kpc. The HI distributions of the two galaxies are offset from the galaxy centers in the direction coincident with the radio continuum tails (Dickey and Gavazzi 1991). All observations are consistent with a scenario where both galaxies are undergoing ram pressure stripping. The two galaxies are located at the periphery of a subcluster merging with the main cluster. The X-ray morphology of Abell 1367 is similar to that of the Norma cluster. The intracluster medium associated with the subcluster is significantly hotter than that of the main cluster. Another galaxy in this region, UGC 6697, has an X-ray tail and a sharp edge in the X-ray distribution at the opposite side (Sun and Vikhlinin 2005). Since Gavazzi et al. (2001) proposed a scenario where UGC 6697 is composed of two interacting galaxies and the system shows a prominent stellar tail, a tidal interaction and ram pressure stripping most probably act together.

Virgo Cluster

In the HI imaging survey of Virgo galaxies (VIVA: VLA Imaging of Virgo galaxies in Atomic gas), Chung et al. (2007) found seven spiral galaxies with long HI tails (➲ *Fig. 5-14*). These tail galaxies have the following properties in common in the HI morphology: (i) the HI is extended well beyond the optical disk only on one side; (ii) the tails differ from tidal bridges; i.e., there is no optical counterpart at the tip of the tail down to r ~ 26 mag arcsec^{-2} in the Sloan Digital Sky Survey (SDSS) images; and (iii) the projected length of the gas tail is larger than half of the symmetric part in HI. These galaxies are found in intermediate- to low-density regions (0.6–1 Mpc in projection from M 87). The tails are all pointing roughly away from M 87. At least three systems show HI truncation to within the stellar disk, providing evidence of a gas–gas interaction. A comparison between the estimated (based on (➲ 5.2)) and observed stripping radii suggests that simple ram pressure stripping could have indeed formed the tails in all but two cases. One of these cases is NGC 4654 where ram pressure stripping is facilitated by a tidal interaction (see below). The HI observations of this spiral galaxy sample represents evidence that ram pressure stripping already begins to affect spiral galaxies around the cluster virial radius. At least 25% of the large spiral galaxies in the region between 0.6 and 1 Mpc from the Virgo cluster center seem to be recent arrivals being stripped of gas.

There are four cases of a mixed interaction, i.e., the combination between a tidal interaction and ram pressure stripping, in the Virgo cluster: NGC 4424, NGC 4654, NGC 4254, and NGC 4438. In the following their properties are described in detail. Other galaxies affected by ram pressure stripping are described in ➲ Sect. 8.3.

The broadband optical appearance of *NGC 4424* is peculiar, with shell-like features and banana-shaped isophotes, suggesting a major gravitational disturbance or a merger (Kenney et al. 1996). This galaxy has extremely modest stellar rotation velocities (~30 km s^{-1}), and stars are supported by random motions as far out as it can be measured (1.4 kpc). The ionized gas kinematics in the core are disturbed and possibly counterrotating. Cortése et al. (2006) suggested that the peculiarities of NGC 4424 are the result of an intermediate-mass merger plus ram pressure stripping which created the long one-sided HI tail detected by Chung et al. (2007).

The spiral galaxy *NGC 4654* is located at a projected distance of 3.4° ~ 1 Mpc from the cluster center (➲ *Fig. 5-5*). It is one of the galaxies with a long one-sided HI tail of Chung et al. (2007). The velocity field of the tail does not show rotation (➲ *Fig. 5-15*). This galaxy

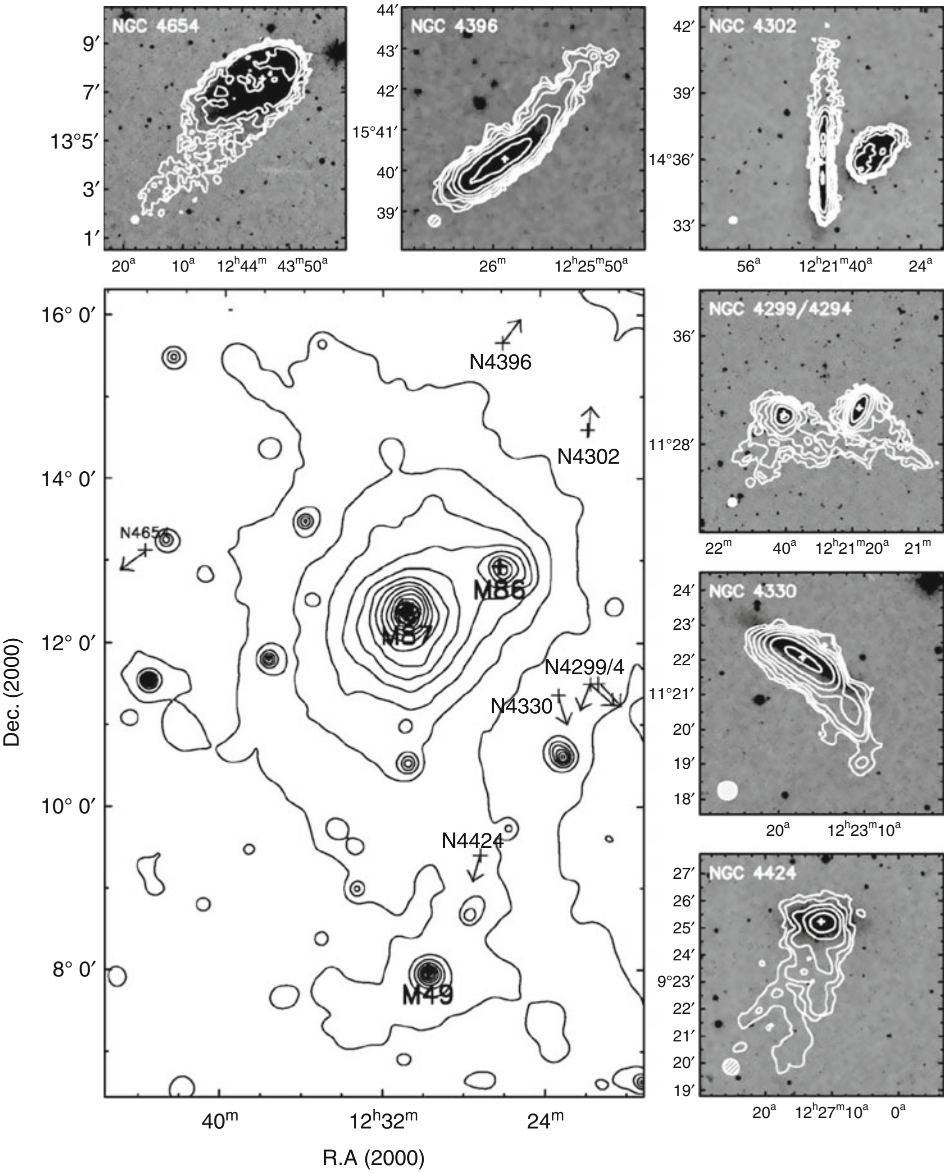


Fig. 5-14

In the main large panel, the locations of the Hι tail galaxies are shown as the crosses on the X-ray background of the Virgo region. The directions of the tails are indicated with the arrows. The second tail of NGC 4299 (east tail) is shown in *light gray*. In the six smaller panels on the top and on the right, we show zoomed views of individual galaxies. The Hι contours (*white*) are shown overlaid on the Digitized Sky Survey (DSS) image in gray scale (From Chung et al. 2007)

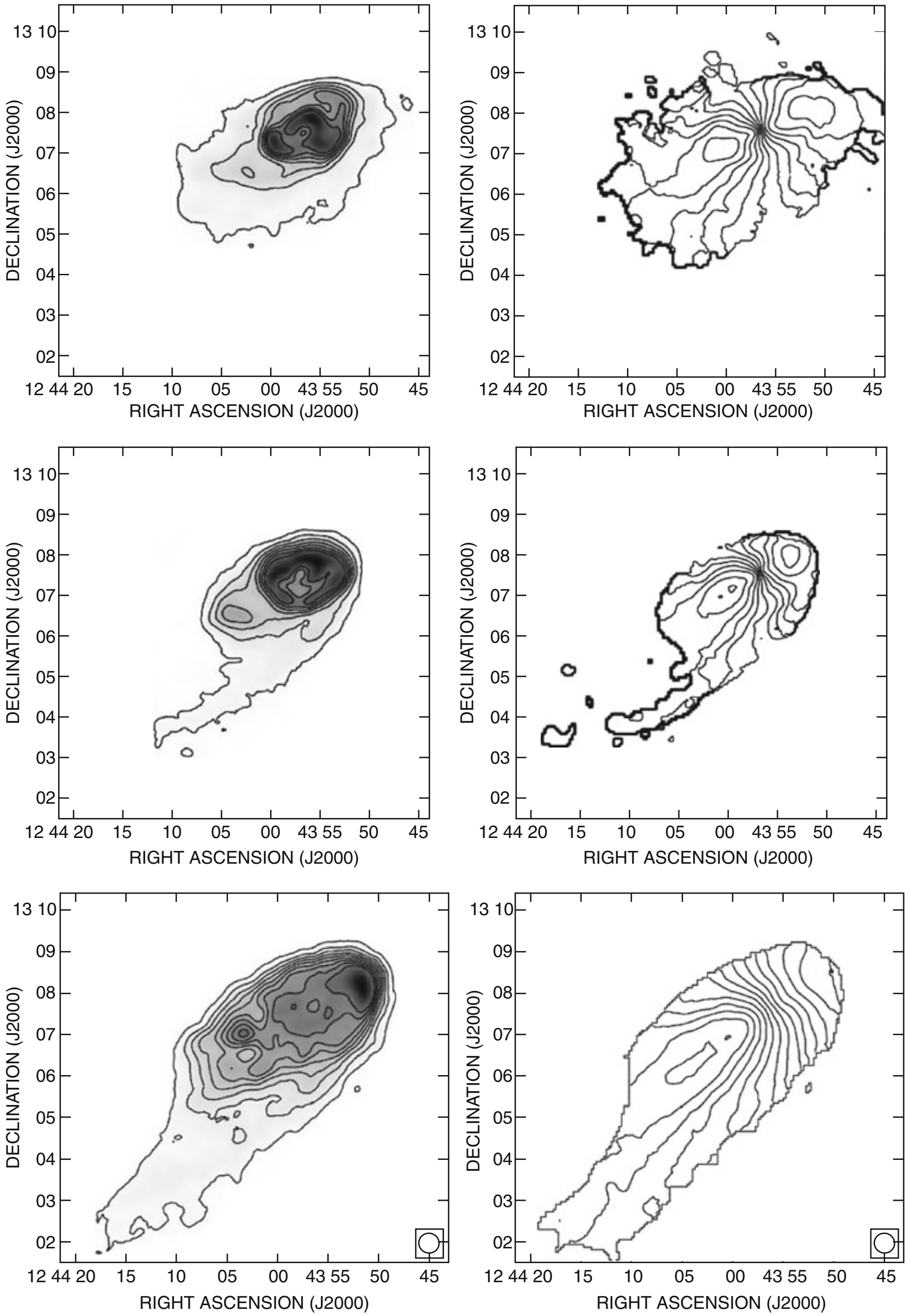


Fig. 5-15

The Virgo cluster spiral NGC 4654. *Left*: HI distribution, *right*: HI velocity field. *Upper panels*: gravitational interaction model, *middle panels*: gravitation interaction + ram pressure stripping model, *lower panels*: HI observations (From Vollmer 2003)

also shows an asymmetric stellar distribution. Numerical simulations using ram pressure as the only perturbation can produce a tail structure of the gas content but cannot account for its kinematical structure. Simulations using a gravitational interaction with the companion galaxy NGC 4639 can account for the asymmetric stellar distribution of NGC 4654 but cannot reproduce the observed extended gas tail. Only a mixed interaction, gravitational and ram pressure, can reproduce all observed properties of NGC 4654 (Vollmer 2003; ➲ *Fig. 5-15*). The tidal interaction loosened the gas from the galaxy's gravitational potential. Only a small amount of ram pressure, about 1/4 of the ram pressure estimated using (➲ 5.2), is necessary to produce the observed HI tail.

The spiral galaxy *NGC 4254* is located at the same projected distance as NGC 4654 (~1 Mpc) on the other side of the Virgo cluster. The spiral structure of NGC 4254 shows an important m = 1 mode, giving it a one-armed appearance. Deep VLA HI observations (Phookun et al. 1993) revealed, in addition to the galactic HI disk component, HI clouds superposed on and beyond the gas disk with velocities that do not follow the disk rotation pattern. Very low surface density atomic gas is found up to ~6′ (~30 kpc) to the north-west of the galaxy center. Subsequent WSRT (Minchin et al. 2007) and Arecibo (Haynes et al. 2008) HI observations revealed the longest low surface density gas tail ever observed with a length of ~250 kpc (➲ *Fig. 5-16*). This tail structure could be reproduced by a rapid (~1,000 km s^{-1}) and close (~60 kpc) flyby of a massive galaxy (1.5 times the mass of NGC 4254) by Duc and Bournaud (2008). The smooth velocity field with a reversal of the gradient in the middle of the tail is a natural result of the dynamical model (➲ *Fig. 5-16*). As in NGC 4654, ram pressure might act on the gas loosened by the gravitational interaction (Vollmer et al. 2005a; Kantharia et al. 2008).

NGC 4438 has the most spectacular tidal tails in the Virgo cluster (➲ *Fig. 5-1*). It is located at a projected distance of only 1° = 300 kpc from the cluster center (M 87). Despite the strong tidal perturbation, ram pressure is the dominant effect on the observed gas distribution and kinematics (Vollmer et al. 2005b). A double line profile is observed in CO(2–1) observations. As in the Coma spiral galaxy NGC 4848, one CO line is due to galactic rotation, whereas the second CO line is most probably due to gas pushed by ram pressure. Recent deep Hα observations of the M 86 region (➲ *Fig. 5-17*; Kenney et al. 2008) revealed a highly complex and disturbed interstellar/intracluster medium. NGC 4438 is connected to M 86 by several faint Hα filaments, which suggests a tidal interaction between the two galaxies, as advocated by Kenney et al. (2008). The timescale and the relative galaxy velocities of the two scenarios (encounter with M 86 or NGC 4435) are about the same. The discovery of an ~190-Myr-old starburst in NGC 4435 by Panuzzo et al. (2007) favors the interaction scenario between NGC 4438 and NGC 4435. Together with NGC 4522 (see below), NGC 4438 is the second galaxy which is undergoing strong ram pressure stripping at a location in the Virgo cluster where this would not be expected from a classical model assuming a spherical, smooth, and static Virgo intracluster medium distribution.

8.2.3 Dwarf Galaxies

Two examples of low-mass irregular galaxies, which have undergone recent ram pressure stripping, were found in the Virgo and Coma cluster. RB 199 is a post-starburst dwarf galaxy with a stellar mass of a few 10^9 $M_\odot$ in the Coma cluster. Yoshida et al. (2008) found a complex of narrow blue filaments, bright blue knots, and Hα-emitting filaments and clouds, which morphologically resembled a complex of "fireballs," extending up to 80 kpc from RB 199. This galaxy

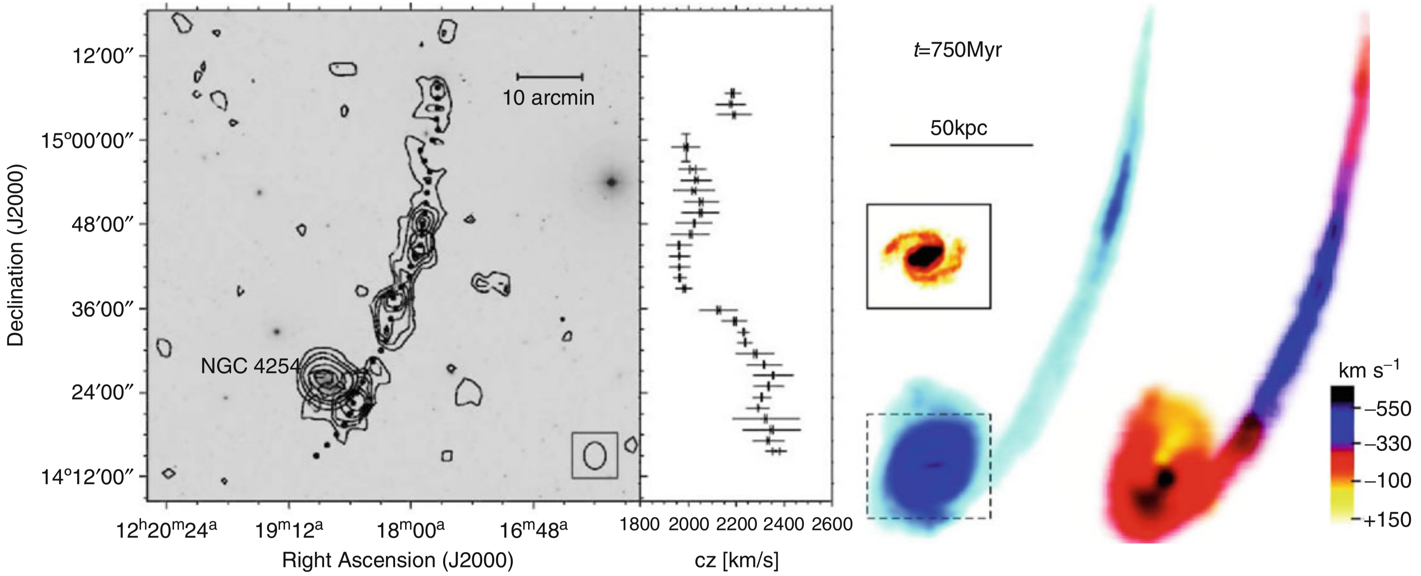


Fig. 5-16

The Virgo cluster spiral NGC 4254. *Left*: Arecibo HI observations (*contour*) on an optical image (from Haynes et al. 2008). *Right*: Dynamical model (From Duc and Bournaud 2008)

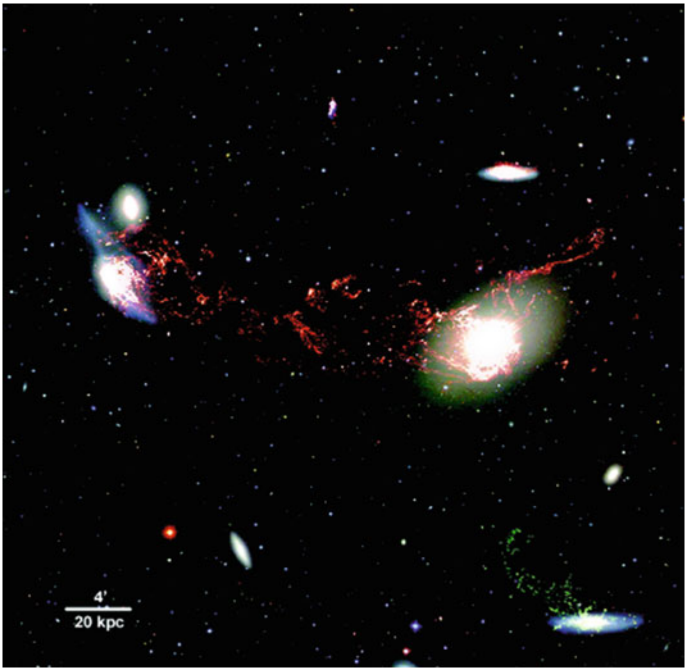


Fig. 5-17

Hα+[N II] image of M 86 region superposed on a color SDSS gri image. The Hα image is stretched to highlight the faint emission. The "low-velocity" (<500 km s^{-1}) Hα+[N II] emission associated with NGC 4438 is colored *red*, and the "high-velocity" (>2,000 km s^{-1}) Hα+[N II] emission associated with NGC 4388 is colored *green* (From Kenney et al. 2008)

has a highly disturbed morphology indicative of a galaxy–galaxy merger remnant (left panel of *Fig. 5-18*). The optical colors of the filaments and knots suggest that most of the stars in the fireballs were formed within several times 100 Myr.

The optical morphology of the Virgo cluster dwarf irregular galaxy IC 3428 ($M_* \sim 10^9\ \mathrm{M}_\odot$) resembles that of RB 199. Hester et al. (2010) found a 17-kpc UV tail of bright knots and diffuse emission behind the galaxy IC 3418 (right panel of *Fig. 5-18*). Hα imaging confirmed that star formation is ongoing in the tail. The stripped gas thus vigorously formed and still forms stars over the last few 100 Myr. This is in contrast to the low star formation rate in stripped gas tails of late-type galaxies (e.g., Vollmer et al. 2008).

8.3 A Holistic View on Ram Pressure Stripping

A first complete ram pressure stripping time sequence was established by Vollmer (2009; *Fig. 5-19*). He combined the results of detailed comparisons between dynamical models and observations of the interstellar medium in ram pressure stripped galaxies in the Virgo cluster. According to his analysis, it is possible to observe ram pressure-induced perturbations

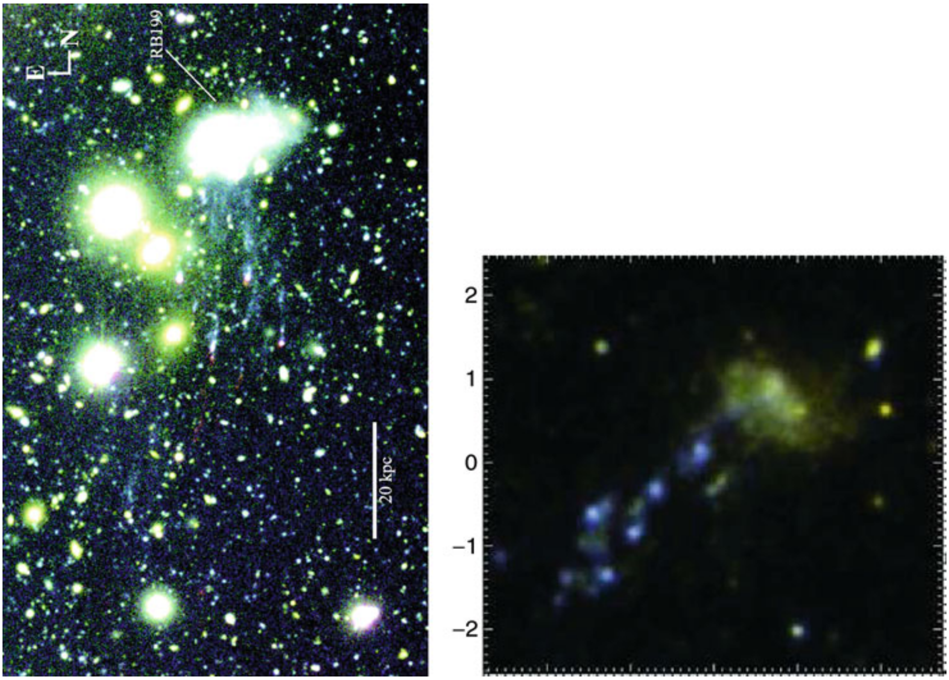


Fig. 5-18
Ram pressure-stripped galaxies with an optical tail structure. *Left panel*: imaging showing a 17-kpc tail of star formation trailing IC 3418 as it plunges through Virgo's intracluster medium. Color composite ultraviolet image: FUV is *blue*, NUV is *red*, and the average UV intensity is *green*. The *y* axis scale is in arcmin (from Hester et al. 2010). *Right panel*: false color (B band: *blue*; R_C band: *green*; Hα: *red*) image of area around RB 199 in the Coma cluster (from Yoshida et al. 2008). A tail structure is visible up to ~80 kpc south to RB 199

~300 Myr around the galaxy's closest approach to the cluster center, i.e., when peak ram pressure occurs if a spherical, smooth, and static intracluster medium distribution is assumed. The relative brevity of this period compared to the galaxy's orbital timescale (several Gyr) explains the rareness (~15%) of observed ram pressure-induced perturbations in Virgo spiral galaxies.

Observationally, the different stages of ram pressure stripping can be recognized in the following way:

- Increasing, moderately strong ram pressure (>50 Myr before peak, class (i)):
  moderately truncated HI disk, extraplanar gas of moderate surface density, continuous velocity field between the disk and the extraplanar region, radio-deficient region, and/or ridge of polarized radio continuum emission at the outer gas disk opposite to the extraplanar region, for example, NGC 4501, NGC 4330.
- Ongoing strong ram pressure (near peak, class (ii)):
  strongly truncated HI disk, extraplanar gas of high surface density, continuous velocity field between the disk and the extraplanar region, radio-deficient region, and/or ridge of

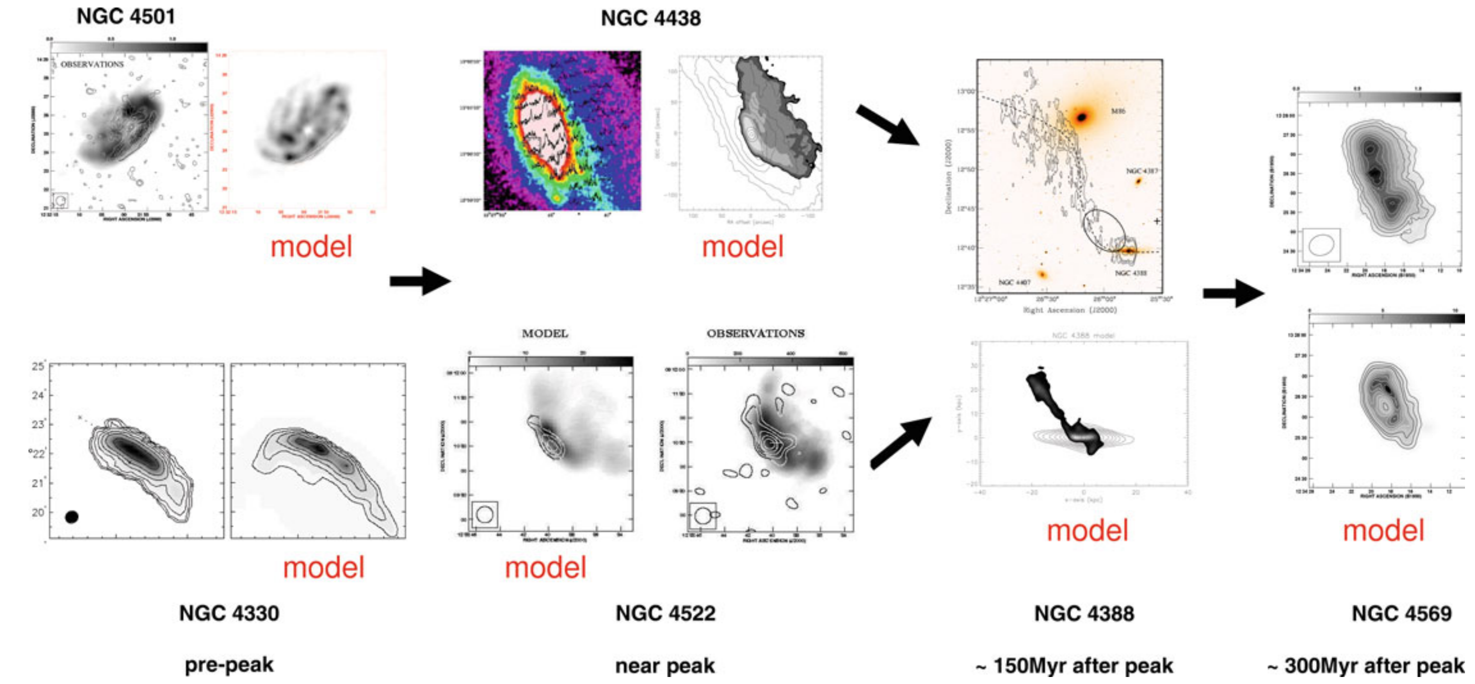


Fig. 5-19
Model-based complete ram pressure stripping time sequence for Virgo cluster spiral galaxies. NGC 4501 (gray scale: H I, contour: polarized radio continuum emission; Vollmer et al. 2008) and NGC 4330 (gray scale and contours: HI; Chung et al. 2007; Vollmer et al. 2012) are approaching the cluster center and are thus in a stage of pre-peak ram pressure (class (i)). NGC 4522 (gray scale: HI, contour: polarized radio continuum emission; Kenney et al. 2004; Vollmer et al. 2006) and NGC 4438 (observations: CO spectra on optical image; model: gray scale: gas surface density; contours: stellar distribution; (Vollmer et al. 2005a) are close to peak ram pressure (class (ii)). NGC 4388 (class (iii); observations: contour: HI Oosterloo and van Gorkom 2005; model: gray scale: gas surface density; contour: stellar distribution; Vollmer and Huchtmeier 2003) and NGC 4569 (class (iv); gray scale and contours: HI; Vollmer et al. 2004a) are leaving the cluster center (From Vollmer 2009)

polarized radio continuum emission at the outer gas disk opposite to the extraplanar region, for example, NGC 4438, NGC 4522.

- Decreasing ram pressure (<200 Myr after peak, class (iii)):
  strongly truncated HI disk, extended extraplanar gas of low surface density, continuous velocity field between the disk and the extraplanar region, and ridge of polarized radio continuum emission at the outer gas disk opposite to the extraplanar region, for example, NGC 4388.
- Decreasing ram pressure (>200 Myr after peak, class (iv)):
  strongly truncated HI disk, perturbed outer gas arms, discontinuous velocity field between the disk and the extraplanar region, ridge of polarized radio continuum emission at the outer gas disk opposite to the extraplanar region, and possible ridge of polarized radio continuum emission at the outer gas disk due to shear motions from the resettling gas, for example, NGC 4569.

The dynamical models yield the 3D velocity vector of the galaxies, the peak ram pressures, and the times to peak ram pressure. In the case of a smooth, static, and spherical intracluster medium, peak ram pressure occurs during the galaxy's closest approach to the cluster center, i.e., when the galaxy's velocity vector is perpendicular to its distance vector. Under these conditions, the galaxy's present line-of-sight distance and its 3D position during peak ram pressure can be calculated. The linear orbital segments derived in this way together with the intracluster medium density distribution from Schindler et al. (1999) are consistent within a factor of 2 with the dynamical simulations for NGC 4501, NGC 4330, and NGC 4569. The impact parameters of the galaxy orbits vary between 200 kpc and 600 kpc.

The analysis of multiwavelength photometry and optical spectra of the gas-free outer regions of the galactic disks allow to derive the time since star formation has been quenched by gas removal via ram pressure stripping. Boselli et al. (2006) used multiband photometry and stellar population modeling to constrain the quenching time. Crowl and Kenney (2008) analyzed the stellar populations in these outer disks of 10 Virgo cluster spiral galaxies, using integral field spectroscopy and UV photometry. All of the galaxies with spatially truncated star formation have outer-disk stellar populations consistent with star formation ending within the last 500 Myr. For approximately half of the galaxies, the truncation ages are consistent with galaxies being stripped in or near the cluster core, where simple ram pressure estimates can explain the observed stripping radius. However, the other half of the galaxies were clearly stripped outside the cluster core. Pappalardo et al. (2010) refined the truncation age determination for NGC 4388 using a nonparametric inversion tool to reconstruct the star formation history of a galaxy from deep VLT spectroscopy and multiband photometry. For all three galaxies where star formation truncation ages based on dynamical models exist (NGC 4569, NGC 4388, NGC 4522), they agree with the timescales derived from spectrophotometry.

Finally, the small edge-on galaxy NGC 4522 (➲ *Fig. 5-2*) deserves special attention. Its gas disk is strongly truncated and extraplanar high column density gas is present (➲ *Fig. 5-2*). Dynamical modeling, the presence of an asymmetric ridge of polarized emission (Vollmer et al. 2004b), and the analysis of the optical spectrum of the gas-free outer disk (Crowl and Kenney 2008) infer strong ongoing ram pressure stripping. However, the galaxy is located at a projected distance of 3.3° ∼ 1 Mpc from the cluster center. Assuming a smooth and static intracluster medium, the calculated ram pressure at this location (➲ 5.2) appears inadequate by an order of magnitude to cause the observed stripping. The most probable scenario is a dynamic, shock-filled intracluster medium with bulk motions and local density enhancements associated with the galaxy group around the elliptical galaxy M 49, which falls into the Virgo cluster from the south. Together with NGC 4438 (see ➋ Sect. 8.2.2), NGC 4522 represents a second galaxy which is not classically stripped in the core of the Virgo cluster.

8.3.1 The Response of the Multiphase ISM and Star Formation to Ram Pressure

As described in ➋ Sect. 5, the interstellar medium consists of different phases. The warm neutral phase is observable in HI line emission and the molecular phase in CO line emission. Hα emission traces the dense warm ionized medium (diffuse gas or HII regions) and diffuse X-ray emission the hot ionized gas phase. In addition, synchrotron emission from cosmic ray gas and the magnetic field is detected in the radio continuum.

Tidal interactions act in the same way on all gas phases and on the stellar component. Ram pressure only affects the ISM. The acceleration of gas clumps by ram pressure depends on their surface density,

$$a = \frac{p_{\text{ram}}}{\Sigma_{\text{ISM}}} = \frac{\rho_{\text{ICM}} v_{\text{gal}}^2}{\Sigma_{\text{ISM}}}, \tag{5.11}$$

where p_{ramp} is the ram pressure, ρ_{ICM} the intracluster medium density, v_{gal} the galaxy velocity with respect to the intracluster medium, and Σ_{ISM} the ISM surface density. Clouds with higher surface densities are thus less affected by ram pressure than clouds with low surface densities. There is evidence that this is actually the case. The edge-on Virgo spiral galaxy NGC 4402 (➋ *Fig. 5-20*) shows signs of ongoing ram pressure stripping (Crowl et al. 2005): the HI disk is strongly truncated and a short HI tail is observed. Moreover, the radio continuum halo is compressed on one side of the galactic disk. Deep optical images show a remarkable dust lane morphology: at half the optical radius, the dust lane of the galaxy curves up and out of the disk. On the leading eastern edge of the interaction, the HI contours appear to cut off inside the dust distribution, suggesting that the less dense gas in this part of the galaxy has already been stripped. To the south of the galactic disk, where the galaxy is relatively clean of gas and dust, there are 1-kpc-long linear dust filaments with a position angle that matches the shape of the radio continuum halo. One of the observed dust filaments has an HII region at its head. These dust filaments are interpreted as large, dense clouds that were initially left behind as the low-density interstellar medium was stripped but were then ablated by the ram pressure wind. The same phenomenon of massive dense molecular clouds which decouple from the ram pressure wind is observed in NGC 4522 and NGC 4438. Both galaxies undergo strong active ram pressure stripping. In both galaxies, small amounts of molecular gas (~10^7 $M_\odot$) with very narrow linewidths are observed in HI-free regions where the bulk of the gas has been removed by ram pressure: in the northern tail of NGC 4438 (Vollmer et al. 2005b) and in NGC 4522 beyond the gas truncation radius (Vollmer et al. 2008). There is one indication that the diffuse warm ionized gas is stripped more efficiently than molecular gas: in the region where extraplanar gas is detected in NGC 4438, the radial velocities of the diffuse Hα emission are significantly offset from ($\Delta v \geq 40$ km s^{-1}), whereas the velocities of the HII regions follow those of the molecular gas (Vollmer et al. 2009).

The evidence of extraplanar star formation is rare in ram pressure stripped Virgo spiral galaxies. Some extraplanar HII regions are detected in Virgo spiral galaxies with HI tails: NGC 4402 (Cortése et al. 2004), NGC 4438 (Kenney et al. 1995), NGC 4388

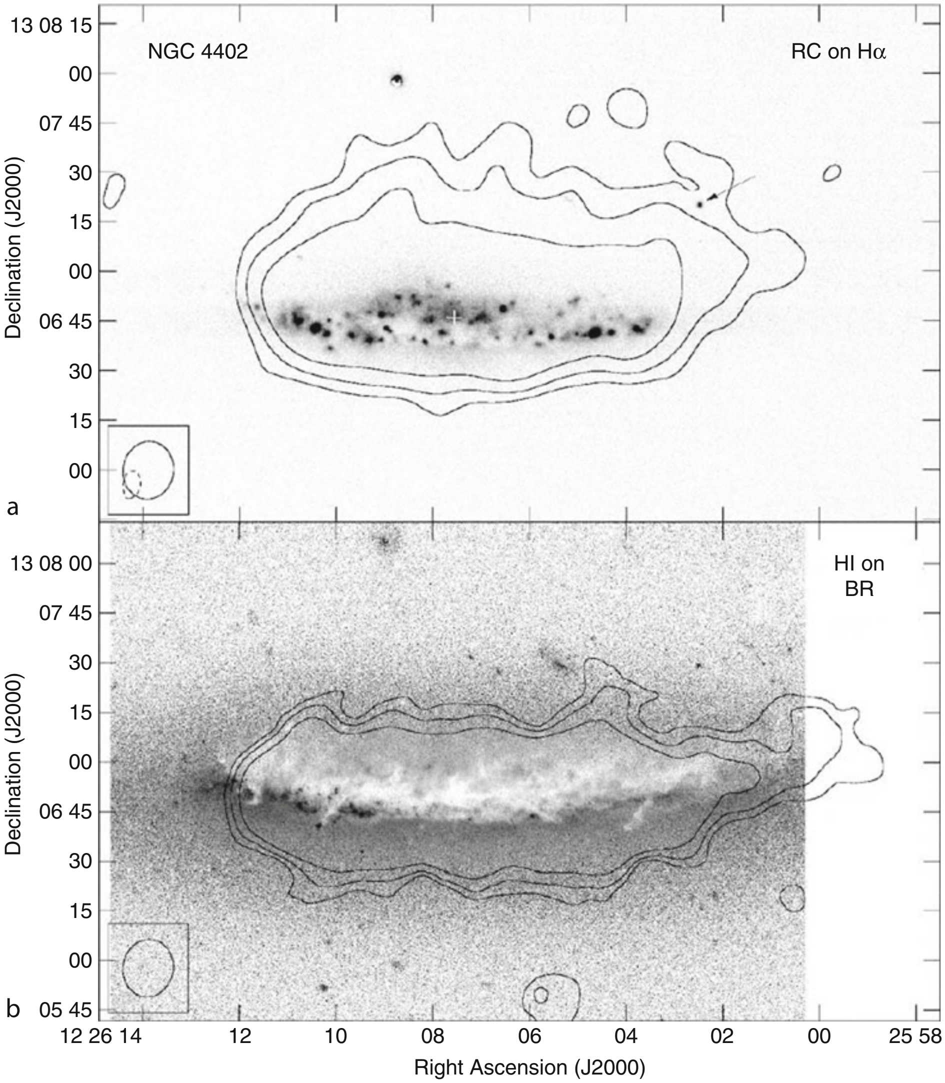


Fig. 5-20

The Virgo cluster spiral galaxy NGC 4402. *Upper panel*: H*α* image together with the outer three radio continuum contours. The arrow marks an extraplanar HII region discovered by Cortése et al. (2004). *Lower panel*: B – R image showing the distribution of dust lanes in the galaxy, along with the outer three HI contours. The kpc dust lanes stick out from the southern edge of the galactic disk and run to the southeast (From Crowl et al. 2005)

(Yoshida et al. 2004), NGC 4522 (Kenney et al. 2004), and NGC 4330 (Abramson et al. 2011). Based on their dynamical model, Vollmer et al. (2008) suggested the following ram pressure stripping scenario: a significant part of the gas is stripped in the form of overdense armlike structures. Molecules and stars form within this dense gas according to the same laws as in

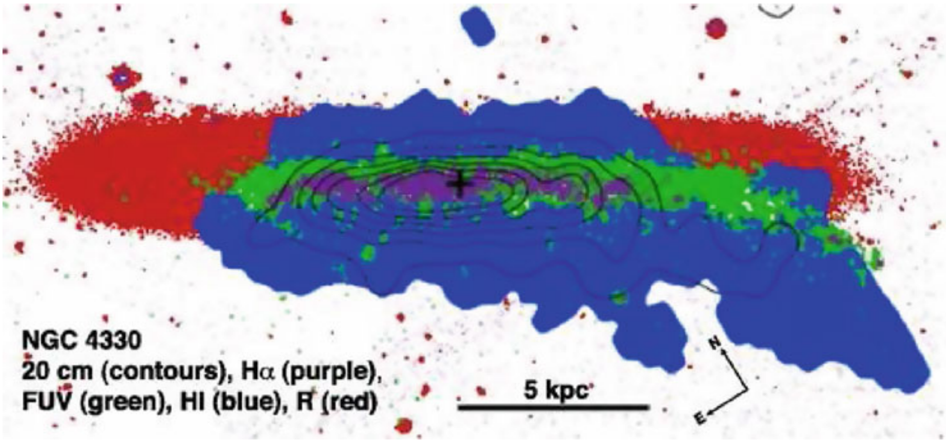


Fig. 5-21
NGC 4330 schematic showing the distributions of R (*red*), FUV (*green*), H*α* (purple), HI (*blue*), and 20-cm radio continuum (*contours*). The radio continuum and HI tails are displaced downwind (SE) of the UV and H*α* tails, indicating that the ISM in this area has been pushed downwind over time (From Abramson et al. 2011)

the galactic disk, i.e., they mainly depend on the total large-scale gas density. Star formation proceeds where the local large-scale gas density is highest. In the absence of a confining gravitational potential, the stripped gas arms will most probably disperse; i.e., the density of the gas will decrease and star formation will cease. This might have happened in NGC 4330 where Abramson et al. (2011) discovered an offset between the UV and HI tails (*Fig. 5-21*). This offset can be explained with the following scenario: since collapsing and star-forming gas clouds decouple from the ram pressure wind, the UV-emitting young stars have the angular momentum of the gas at the time of their creation. On the other hand, the gas is constantly pushed by ram pressure and its density decreases. At a certain density threshold, star formation ceases in the displaced gas. The observed gas tails has a very low present star formation efficiency. The UV emission indicates that the past star formation efficiency (~100 Myr) was much higher.

9 The Detailed Picture

It is difficult to find cluster spiral galaxies with clear signs of ongoing ram pressure stripping and unambiguous examples are rare. The signposts of ram pressure stripping are (i) a symmetric old stellar disk together with a gas disk truncated inside the optical radius, (ii) a one-sided gas tail, and (iii) a radio-deficient region and/or an asymmetric polarized radio continuum ridge located in the outer gas disk opposite to the gas tail. In the absence of a long gas tail, it is easier to recognize extraplanar gas in edge-on galaxies. Therefore, we are somewhat biased against this particular projection. Although the distortions of the gas distribution can be quite dramatic, the gas velocity field is still smooth and regular. In the Virgo cluster, these asymmetries can be observed during ~300 Myr, compared to an orbital timescale of a few Gyr. This partly explains the rareness of published cases with ongoing ram pressure stripping. In the classical picture, a galaxy approaches the cluster center on a highly eccentric orbit. The intracluster medium is in hydrostatic equilibrium and does not move. Ram pressure stripping thus occurs close to the cluster core when (➤ 5.2) is fulfilled. The distance from the cluster center where a significant amount of gas is removed from the galaxy depends on the radial distribution of intracluster medium which can be very different from cluster to cluster (e.g., Virgo and Coma). We have learned from the Virgo spiral galaxy NGC 4522 that ram pressure stripping can occur far away from the cluster center in regions where classical ram pressure by a static intracluster medium is insufficient. The explanation is that the intracluster medium is moving at a high speed in the opposite direction of the galaxy's motion. Since ram pressure depends on the square of the intracluster medium velocity with respect to galaxy, this can dramatically enhance ram pressure. Thus, whenever the intergalactic medium of a subcluster or galaxy group collides with the intracluster medium of the main cluster during a cluster–subcluster or a cluster–group merger, ram pressure stripping can occur at the cluster periphery provided that there are galaxies which move against the direction of the infalling intergalactic medium.

The detailed comparison between dynamical models and gas distributions and velocity fields of Virgo spiral galaxies showed that the Gunn and Gott critrion (➤ 5.2) is valid within a factor of 2 (Vollmer 2009). The models of different types (3D hydro, SPH, sticky particles) agree in this respect (see ➤ Sect. 4). It is not enough to reproduce the observed gas distribution, because the results can be ambiguous. The observed gas velocity field adds important constraints for the modeling. The motions in the plane of the sky can be probed via the polarized radio continuum emission which is sensitive to compression and shear. The main ingredients of ram pressure stripping are (i) ram pressure; (ii) galactic rotation; (iii) gas shadowing, i.e., gas on the windward side can shadow gas located on the downwind side; and (iv) for long gas tails intracluster medium – ISM mixing. Mixing requires Kelvin–Helmholtz instabilities and can thus only be simulated by 3D Eulerian hydrodynamics.

Most of the gas tails are observed in HI and only a few in X-rays, Hα, and radio continuum emission. Unfortunately, there is only one simultaneous detection of a gas tail in HI and X-rays, NGC 4848 in the Coma cluster. Why do some tails are X-ray bright and form more stars (ESO 137-001 in Abell 3627) than others (NGC 4388 in Virgo) which are fainter in X-rays? Tonnesen et al. (2011) suggested that the primary requirement is a high-pressure intracluster medium. This is because the stripped tail is mostly in pressure equilibrium with the intracluster medium, but mixing leaves it with densities and temperatures intermediate between the cold gas in the disk and the hot intracluster medium. Given a high enough intracluster medium pressure, the mixed gas lies in the X-ray bright region of the phase diagram. This suggestion is consistent with the higher intracluster medium temperature and thus pressure of Abell 3627. This also explains why the X-ray bight tail of ESO 137-001 shows diffuse Hα emission.

The work on NGC 4522 and NGC 4330 suggests that the gas is stripped in an arm structure with high densities. As long as the gas is of high density, it forms stars. Once the gas is pushed out of the galactic disk, the gravitational confinement vanishes and the bulk of the gas expands and is mixed into the tenuous hot intracluster medium. Since the stripping timescale is long compared to the lifetime of a molecular cloud, the ISM is stripped as an entity. Only the largest molecular cloud complexes might decouple from the ram pressure wind and stay behind (NGC 4522, ➤ *Fig. 5-2*; NGC 4402, ➤ *Fig. 5-20*, NGC 4438 in the Virgo cluster). There are indications for a modest enhancement of the star formation rate per surface area in NGC 4522 (Crowl and Kenney 2006). At the same time, the total star formation rate of the galaxy is reduced because of the truncation of the gas and star-forming disk. Thus, whereas the local star formation rate can be enhanced up to a factor of a few during a ram pressure stripping event, the total star formation rate decreases in most cases.

The gas stripping of gas-rich dwarf irregular galaxies proceeds in a different way. The star formation efficiency of the stripped gas is higher leading to stellar tails which are observable in the UV and optical domain (RB 199 in Coma and IC 3418 in Virgo).

Gravitational interactions, like close flybys of massive galaxies, can loosen the gas from the galaxy's gravitational potential, making ram pressure stripping more efficient (NGC 4654, ➲ *Fig. 5-15*; NGC 4438 in the Virgo cluster). In an extreme case, a gravitational interaction might even be responsible for a 100-kpc gas tail (NGC 4254 in the Virgo cluster; ➲ *Fig. 5-16*).

The end-product of ram pressure stripping are spiral galaxies with a truncated gas and star-forming disk. These stripped late-type galaxies can easily be misclassified as early-type spirals (Koopmann and Kenney 1998). Star formation will then consume the gas, the gas surface density and star formation rate will decrease, and the galaxy will become passive or anemic (NGC 4548, NGC 4579, NGC 4569 in the Virgo cluster). After the stripping, the outer gas-free disk has a post-starburst spectrum with strong Balmer absorption lines and no significant emission from ongoing star formation (Crowl and Kenney 2006).

For the transformation of the spiral galaxies into lenticulars, a morphological transformation leading to larger bulges seems to be necessary as discussed in ➲ Sect. 7. Tidal interactions are the prime candidate for this transformation which happened either via slow encounters in a group environment before infall into the galaxy cluster or via harassment or single close flybys of massive galaxies in the cluster environment. An example for an ongoing spiral–S0 transformation at the cluster periphery is the Virgo cluster galaxy NGC 4438. A single gravitational interaction leads to important stellar tidal tails (➲ *Fig. 5-17*). Once the galaxy has passed the cluster core, these tails will be stripped by the gravitational potential of the cluster and NGC 4438 will resemble an S0 galaxy. The tidally stripped stars will contribute to the diffuse intracluster light (see, e.g., Mihos et al. 2005).

10 A Local View on the Butcher–Oemler Effect

Butcher and Oemler (1984) were the first to report an increasing fraction of blue galaxies in 33 rich galaxy clusters out to a redshift of $z \sim 0.5$, Subsequent studies showed similar trends for the star formation rate (e.g., Balogh et al. 1999; Poggianti et al. 2006) and galaxy morphology (e.g., Dressler et al. 1997; Treu et al. 2003), or spectral properties (Ellingson et al. 2001). The fraction of blue spirals is determined by the cluster richness and redshift (Margoniner et al. 2001). This global effect can be understood in a scenario where the rate of field galaxy infall onto clusters decreases with decreasing redshift (Kauffmann 1995; Abraham et al. 1996; Ellingson et al. 2001). Since the mean cluster mass increases with decreasing redshift, the specific infall rate per unit cluster mass decreases at an even faster rate. Furthermore, because of the sharply declining global star formation rate as a function of decreasing redshift (Hopkins and Beacom 2006), field galaxies form on average more stars and are bluer at higher redshifts. A more detailed view of the Butcher–Oemler effect is that of a significant evolution in the fractional population gradient of early- and late-type galaxies (Ellingson et al. 2001). Both low- and high-redshift clusters have similar populations in the cluster cores, but higher redshift clusters have steeper gradients and more star-forming galaxies at radii outside of the core region. The blue galaxy fraction at a given redshift depends on cluster mass (richness, Margoniner et al. 2001; X-ray temperature Urquhart et al. 2010), but its gradient with respect to redshift is approximatelly the same for all clusters (Margoniner et al. 2001). The Buther–Oemler effect thus has an environmental and a cosmological component. The main effect of the cluster environment is to quench star formation of the infalling galaxies, and more massive galaxy clusters are more efficient in stopping the star formation of their galaxies. The cause of the environmental component of the Butcher–Oemler effect (ram pressure stripping or tidal interactions) has still to be determined. It is well possible that both effects are equally needed. The local view of the cosmological component of the Butcher–Oemler effect is that the fraction of dynamically young clusters as the Virgo, Abell 1367, and Norma clusters with a high infall rate of blue field galaxies increases with redshift.

The studies of local galaxy clusters teach us that cluster environments can be very different. In a tentative sequence for cluster evolution, a galaxy cluster begins as a Fornax-like cluster with a high fraction of healthy spirals and a small amount of hot intracluster medium peaking on the central luminous galaxy. In this environment, gravitational interactions dominate and ram pressure stripping does not play a role for galaxy evolution within the cluster. When important groups of galaxies fall into the cluster, it will look like the Virgo cluster with a spiral-rich galaxy distribution which shows a lot of substructure. If the infalling groups contain a significant amount of hot intragroup gas, the intracluster–intragroup gas merging will heat the intracluster medium and lead to important bulk motions. In these highly dynamic environment with increasing intracluster medium mass, ram pressure becomes more and more efficient. Ram pressure typically begins to affect the galaxies' outer gas disks inside the cluster's Virial radius. The clustercentric distance from which ram pressure significantly reduces the galaxies' gas and star-forming disks depends on (i) the central density and core radius of the intracluster medium distribution and (ii) the intracluster medium bulk motions. Only galaxies on eccentric orbits, leading them deeply into the cluster core or galaxies flying against bulk motions will be stripped efficiently. The gas stripping is then quasi-instantaneous. For the other galaxies on less eccentric orbits close flybys of massive perturbers can loosen the gas in the outer parts of the disk which can then be stripped by a relatively small amount of ram pressure. The timescale for this mixed interaction is longer and depends on the probability of a close gravitational interaction. Galaxy harassment, or multiple close flybys, mainly changes the morphology of low-mass galaxies. In addition, spiral galaxies consume their gas via star formation without a supply of fresh gas. The galaxy cluster grows steadily in this way.

If another smaller galaxy cluster with a significant amount of intracluster gas hits the main cluster, the intracluster medium temperature and bulk motions increase. With increasing mass and temperature of the intracluster medium, its central density and core radius increase (Jones and Forman 1999; Schindler et al. 1999). We then have a cluster-like the Norma or Abell 1367 clusters where ram pressure stripping is enhanced in the gas merger region. When the intracluster medium is relaxed again, the cluster will look like the Coma cluster. Through the dynamical buildup of the intracluster medium, ram pressure becomes more and more important for the evolution of the star formation of cluster galaxies. Ram pressure stripped spirals consume the residual gas and the spiral galaxy becomes passive or anemic. The end products might be relatively faint lenticular galaxies.

Global morphological transformation of spiral galaxies into lenticulars within the cluster environment certainly needs important tidal interactions. It should not be forgotten that some galaxies are already preprocessed within infalling groups, where they have already lost a fraction of their ISM and might have changed morphology. The formation of lenticular galaxies is thus a heterogeneous process which acts in group and cluster environments (see also Moran et al. 2007; van den Bergh 2009).

11 Conclusions and Outlook

We have seen that the environment modifies the properties of a galaxy. The primary effects are gas removal, quenching of star formation, and a possible morphological transformation. Those galaxies that live in a group undergo predominantly slow tidal interactions with other galaxies. Because of their high galaxy density, compact groups show dramatic tidal tails and galaxy mergers. In rare cases, ram pressure can play a role, if a galaxy flys through a gaseous tidal tail or if the group contains a considerable amount of intragroup gas. The end product of a galaxy merger is a lenticular or elliptical galaxy. Tidal interactions will drive the ISM of the inner disk into the galaxy center and loosen the ISM of the outer disk. Star formation then consumes the ISM in the galactic disk. In this way, some group galaxies can become HI deficient and/or undergo morphological transformation.

A significant fraction of galaxies are preprocessed in a group environment before falling into a cluster. In galaxy clusters, tidal interactions are fast and numerous (galaxy harassment). These interactions will mainly modify the morphology of low-mass galaxies. If the intracluster gas mass is high and the core radius of its distribution is large, or if the intracluster medium is moving due to a cluster–cluster merger, ram pressure stripping becomes an important agent of galaxy evolution. Ram pressure stripping is most efficient for galaxies on highly eccentric orbits and for galaxies moving against bulk motions of the intracluster gas. The galaxies' ISM is then removed instantaneously according to the Gunn and Gott criterion (➤ 5.2), and the galaxies become HI deficient. There is now evidence in local galaxy clusters where and how ram pressure acts on a spiral galaxy. Tidal interactions and ram pressure stripping can act simultaneously, enhancing the ISM removal. These mixed interactions occur preferentially at the cluster periphery (≥1 Mpc). Signs of ram pressure stripping can be detected up to about one virial radius of the cluster. The time window, in which we can unambiguously identify the effects of ram pressure stripping, is small compared to the orbital period. Therefore, it is not easy to find these galaxies in a cluster. Large imaging HI surveys are particular useful to identify ram pressure stripping candidates. Polarized radio continuum observations can be used to verify the ram pressure stripping hypothesis. The physics of the interaction can then be studied by multiwavelength observations from the X-ray to radio domain. For a limited sample of galaxies, it is now possible to constrain their 3D orbits and interaction histories within the cluster. We are now at the point where we can study the reaction (phase change, star formation) of the multiphase ISM (molecular, atomic, ionized) to ram pressure.

For the moment, detailed interaction diagnostics are limited to local galaxy clusters where the resolution of current telescopes is sufficiently high. The HI imaging resolution for Coma galaxies is already too coarse to permit detailed modeling. Furthermore, the most distant galaxy clusters where galaxies could be detected (not resolved) in HI emission are at $z = 0.2$ (Abell 963, Abell 2192; Verheijen et al. 2007). The main obstacles to observing HI in distant galaxies are the necessarily long integration times and man-made interference outside the protected 21-cm band. The square kilometer array (SKA) will entirely change this situation, allowing us detailed HI line and radio continuum imaging of cluster galaxies in local clusters beyond Virgo and giving us access to the gas content of galaxies at higher redshifts. With ALMA, it will be possible to study the reaction of the dense gas phase to ram pressure stripping (the decoupling of giant molecular clouds from the ram pressure wind) and investigate the total molecular gas content of spiral galaxies in clusters beyond the local universe. LOFAR will give us access to the population of older cosmic ray electrons. This will greatly increase our knowledge on the effects of ram pressure stripping on the cosmic ray gas. The Herschel satellite has already and will improve our understanding of the role of dust in environmental interactions. The road to a better understanding of environmental effects on galaxy evolution is thus lined with beautiful upcoming instruments.

Acknowledgments

I would like to thank A. Boselli and J. Köppen for their careful reading of the manuscript and useful remarks and discussions.

References

Abadi, M. G., Moore, B., & Bower, R. G. 1999, MNRAS, 308, 947

Abraham, R. G. et al. 1996, ApJ, 471, 694

Abramson, A., Kenney, J. D. P., Crowl, H. H., Chung, A., van Gorkom, J. H., Vollmer, B., & Schiminovich, D. 2011, AJ, 141, 164

Allam, S., Assendorp, R., Longo, G., Braun, M., & Richter, G. 1996, A&AS, 117, 39

Andersen, V., & Owen, F. N. 1995, AJ, 109, 1582

Balogh, M. L., Morris, S. L., Yee, H. K. C., Carlberg, R. G., & Ellingson, E. 1999, ApJ, 527, 54

Balogh, M. L., Baldry, I. K., Nichol, R., Miller, C., Bower, R., & Glazebrook, K. 2004, ApJ, 615, L101

Balsara, D., Livio, M., & O'Dea, C. P. 1994, ApJ, 437, 83

Bamford, S. P., Rojas, A. L., Nichol, R. C., Miller, C. J., Wasserman, L., Genovese, C. R., & Freeman, P. E. 2008, MNRAS, 391, 607

Barnes, J. E., & Hernquist, L. 1996, ApJ, 471, 115

Barway, S., Wadadekar, Y., Kembhavi, A. K., & Mayya, Y. D. 2009, MNRAS, 394, 1991

Beck, R. 2001, SSRv, 99, 243

Beck, R. 2005, LNP, 664, 41

Bekki, K., Couch, W. J., & Drinkwater, M. J. 2001, ApJ, 552, L105

Bekki, K., Couch, W. J., & Shioya, Y. 2002, ApJ, 577, 651

Bicay, M. D., & Giovanelli, R. 1987, ApJ, 321, 645

Bigiel, F., Leroy, A., Walter, F., Brinks, E., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2846

Biller, B. A., Jones, C., Forman, W. R., Kraft, R., & Ensslin, T. 2004, ApJ, 613, 238

Binggeli, B., Sandage, A., & Tammann, G. A. 1988, ARA&A, 26, 509

Binggeli, B., Tarenghi, M., & Sandage, A. 1990, A&A, 228, 42

Biviano, A. 1998, Untangling Coma Berenices: a new vision of an old cluster, in Proceedings of the Meeting held in Marseilles (France), June 17–20, 1997, ed. A. Mazure, F. Casoli, F. Durret, & D. Gerbal (Word Scientific Publishing Co Pte Ltd, Singapore), 1

Blanton, M. R., & Moustakas, J. 2009, ARA&A, 47, 159

Böhringer, H., Briel, U. G., Schwarz, R. A., Voges, W., Hartner, G., & Trümper, J. 1994, Nature, 368, 828

Böhringer, H., Neumann, D. M., Schindler, S., & Kraan-Korteweg, R. C. 1996, ApJ, 467, 168

Boselli, A., & Gavazzi, G. 2006, PASP, 118, 517

Boselli, A., Gavazzi, G., Lequeux, J., Buat, V., Casoli, F., Dickey, J., & Donas, J. 1995, A&A, 300, L13

Boselli, A., Mendes de Oliveira, C., Balkowski, C., Cayatte, V., & Casoli, F. 1996, A&A, 314, 738

Boselli, A., Lequeux, J., & Gavazzi, G. 2002, A&A, 384, 33

Boselli, A., Boissier, S., Cortese, L., Gil de Paz, A., Seibert, M., Madore, B. F., Buat, V., & Martin, D. C. 2006, ApJ, 651, 811

Boselli, A., Boissier, S., Cortese, L., & Gavazzi, G. 2008, ApJ, 674, 74

Boulares, A., & Cox, D. P. 1990, ApJ, 365, 544

Bravo-Alfaro, H., Cayatte, V., van Gorkom, J. H., & Balkowski, C. 2000, AJ, 119, 580

Briel, U. G., Henry, J. P., & Boehringer, H. 1992, A&A, 259, L31

Broeils, A. H., & Rhee, M.-H. 1997, A&A, 324, 877

Butcher, H., & Oemler, A., Jr. 1978, ApJ, 226, 559

Butcher, H., & Oemler, A., Jr. 1984, ApJ, 285, 426

Byrd, G., & Valtonen, M. 1990, ApJ, 350, 89

Casoli, F., Boisse, P., Combes, F., & Dupraz, C. 1991, A&A, 249, 359

Casoli, F., Dickey, J., Kazes, I., Boselli, A., Gavazzi, P., & Baumgardt, K. 1996, A&A, 309, 43

Cayatte, V., van Gorkom, J. H., Balkowski, C., & Kotanyi, C. 1990, AJ, 100, 604
Cayatte, V., Kotanyi, C., Balkowski, C., & van Gorkom, J. H. 1994, AJ, 107, 1003
Chamaraux, P., Balkowski, C., & Gerard, E. 1980, A&A, 83, 38
Chilingarian, I., Cayatte, V., Revaz, Y., Dodonov, S., Durand, D., Durret, F., Micol, A., & Slezak, E. 2009, Sci, 326, 1379
Chilingarian, I. V., Mieske, S., Hilker, M., & Infante, L. 2011, MNRAS, 412, 1627
Christlein, D., & Zabludoff, A. I. 2004, ApJ, 616, 192
Chung, A., van Gorkom, J. H., Kenney, J. D. P., & Vollmer, B. 2007, ApJ, 659, L115
Chung, A., van Gorkom, J. H., Kenney, J. D. P., Crowl, H., & Vollmer, B. 2009, AJ, 138, 1741
Clemens, M. S., Alexander, P., & Green, D. A. 2000, MNRAS, 312, 236
Combes, F., Dupraz, C., Casoli, F., & Pagani, L. 1988, A&A, 203, L9
Cortés, J. R., Kenney, J. D. P., & Hardy, E. 2006, AJ, 131, 747
Cortése, L., Gavazzi, G., Boselli, A., & Iglesias-Paramo, J. 2004, A&A, 416, 119
Cortése, L., Gavazzi, G., Boselli, A., Franzetti, P., Kennicutt, R. C., O'Neil, K., & Sakai, S. 2006, A&A, 453, 847
Cortése, L., et al. 2010a, A&A, 518, L63
Cortése, L., et al. 2010b, A&A, 518, L49
Côté, P., et al. 2006, ApJS, 165, 57
Croston, J. H., Hardcastle, M. J., & Birkinshaw, M. 2005, MNRAS, 357, 279
Crowl, H. H., & Kenney, J. D. P. 2006, ApJ, 649, L75
Crowl, H. H., & Kenney, J. D. P. 2008, AJ, 136, 1623
Crowl, H. H., Kenney, J. D. P., van Gorkom, J. H., & Vollmer, B. 2005, AJ, 130, 65
Dame, T. M., Hartmann, D., & Thaddeus, P. 2001, ApJ, 547, 792
Davies, R. D., & Lewis, B. M. 1973, MNRAS, 165, 231
Davies, J. I., et al. 2010, A&A, 518, L48
da Rocha, C., & Mendes de Oliveira, C. 2005, MNRAS, 364, 1069
de la Rosa, I. G., de Carvalho, R. R., Vazdekis, A., & Barbuy, B. 2007, AJ, 133, 330
Dickey, J. M., & Gavazzi, G. 1991, ApJ, 373, 347
Domainko, W., et al. 2006, A&A, 452, 795
Donnelly, R. H., Markevitch, M., Forman, W., Jones, C., David, L. P., Churazov, E., & Gilfanov, M. 1998, ApJ, 500, 138
Dressler, A. 1980, ApJ, 236, 351
Dressler, A. 1986, ApJ, 301, 35
Dressler, A. 2004, in Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution, from the Carnegie Observatories Centennial Symposia, Carnegie Observatories Astrophysics Series, ed. J. S. Mulchaey, A. Dressler, & A. Oemler (Cambridge, UK: Cambridge University Press), 206
Dressler, A., & Sandage, A. 1983, ApJ, 265, 664
Dressler, A., et al. 1997, ApJ, 490, 577
Dressler, A., Smail, I., Poggianti, B. M., Butcher, H., Couch, W. J., Ellis, R. S., & Oemler, A., Jr. 1999, ApJS, 122, 51
Drinkwater, M. J., Jones, J. B., Gregg, M. D., & Phillipps, S. 2000, PASA, 17, 227
Drinkwater, M. J., Gregg, M. D., & Colless, M. 2001, ApJ, 548, L139
Duc, P.-A., & Bournaud, F. 2008, ApJ, 673, 787
Duc, P.-A., Bournaud, F., & Masset, F. 2004, A&A, 427, 803
Dunn, L. P., & Jerjen, H. 2006, AJ, 132, 1384
Eke, V. R., Baugh, C. M., Cole, S., Frenk, C. S., King, H. M., & Peacock, J. A. 2005, MNRAS, 362, 1233
Ellingson, E., Lin, H., Yee, H. K. C., & Carlberg, R. G. 2001, ApJ, 547, 609
Fasano, G., Poggianti, B. M., Couch, W. J., Bettoni, D., Kjærgaard, P., & Moles, M. 2000, ApJ, 542, 673
Ferguson, H. C., & Binggeli, B. 1994, A&ARv, 6, 67
Finoguenov, A., Pietsch, W., Aschenbach, B., & Miniati, F. 2004a, A&A, 415, 415
Finoguenov, A., Briel, U. G., Henry, J. P., Gavazzi, G., Iglesias-Paramo, J., & Boselli, A. 2004b, A&A, 419, 47
Forman, W., Schwarz, J., Jones, C., Liller, W., & Fabian, A. C. 1979, ApJ, 234, L27
Fryxell, B., et al. 2000, ApJS, 131, 273
Fumagalli, M., & Gavazzi, G. 2008, A&A, 490, 571
Gaetz, T. J., Salpeter, E. E., & Shaviv, G. 1987, ApJ, 316, 530
Gavazzi, G. 1987, ApJ, 320, 96
Gavazzi, G. 1989, ApJ, 346, 59
Gavazzi, G., & Boselli, A. 1999a, A&A, 343, 86
Gavazzi, G., & Boselli, A. 1999b, A&A, 343, 93
Gavazzi, G., Contursi, A., Carrasco, L., Boselli, A., Kennicutt, R., Scodeggio, M., & Jaffe, W. 1995, A&A, 304, 325
Gavazzi, G., Marcelin, M., Boselli, A., Amram, P., Vílchez, J. M., Iglesias-Paramo, J., & Tarenghi, M. 2001, A&A, 377, 745
Gavazzi, G., Boselli, A., van Driel, W., & O'Neil, K. 2005, A&A, 429, 439
Gavazzi, G., Boselli, A., Cortese, L., Arosio, I., Gallazzi, A., Pedotti, P., & Carrasco, L. 2006a, A&A, 446, 839
Gavazzi, G., O'Neil, K., Boselli, A., & van Driel, W. 2006b, A&A, 449, 929
Gavazzi, G., et al. 2008, A&A, 482, 43
Geha, M., Guhathakurta, P., & van der Marel, R. P. 2003, AJ, 126, 1794

Gerber, R. A., & Lamb, S. A. 1994, ApJ, 431, 604

Ghigna, S., Moore, B., Governato, F., Lake, G., Quinn, T., & Stadel, J. 1998, MNRAS, 300, 146

Giovanelli, R., & Haynes, M. P. 1985, ApJ, 292, 404

Gómez, P. L., et al. 2003, ApJ, 584, 210

Graham, A. W. 2002, ApJ, 568, L13

Grebenev, S. A., Forman, W., Jones, C., & Murray, S. 1995, ApJ, 445, 607

Gunn, J. E., & Gott, J. R., III 1972, ApJ, 176, 1

Hashimoto, Y., Oemler, A., Jr., Lin, H., & Tucker, D. L. 1998, ApJ, 499, 589

Haynes, M. P., & Giovanelli, R. 1984, AJ, 89, 758

Haynes, M., Arber, T. D., & Verwichte, E. 2008, A&A, 479, 235

Haşegan, M., et al. 2005, ApJ, 627, 203

Helou, G., Soifer, B. T., & Rowan-Robinson, M. 1985, ApJ, 298, L7

Henriksen, M., & Byrd, G. 1996, ApJ, 459, 82

Hester, J. A., et al. 2010, ApJ, 716, L14

Hickson, P., Mendes de Oliveira, C., Huchra, J. P., & Palumbo, G. G. 1992, ApJ, 399, 353

Hilker, M., Infante, L., Vieira, G., Kissler-Patig, M., & Richtler, T. 1999, A&AS, 134, 75

Hilker, M., Baumgardt, H., Infante, L., Drinkwater, M., Evstigneeva, E., & Gregg, M. 2007, A&A, 463, 119

Hinz, J. L., Rieke, G. H., & Caldwell, N. 2003, AJ, 126, 2622

Hopkins, A. M., & Beacom, J. F. 2006, ApJ, 651, 142

Horellou, C., Casoli, F., & Dupraz, C. 1995, A&A, 303, 36

Huchtmeier, W. K. 1997, A&A, 325, 473

Iglesias-Páramo, J., & Vílchez, J. M. 1999, ApJ, 518, 94

Jáchym, P., Palouš, J., Köppen, J., & Combes, F. 2007, A&A, 472, 5

Jáchym, P., Köppen, J., Palouš, J., & Combes, F. 2009, A&A, 500, 693

Jaffe, W., & Gavazzi, G. 1986, AJ, 91, 204

Jeltema, T. E., Mulchaey, J. S., Lubin, L. M., & Fassnacht, C. D. 2007, ApJ, 658, 865

Johnson, K. E., Hibbard, J. E., Gallagher, S. C., Charlton, J. C., Hornschemeier, A. E., Jarrett, T. H., & Reines, A. E. 2007, AJ, 134, 1522

Jones, C., & Forman, W. 1999, ApJ, 511, 65

Kantharia, N. G., Rao, A. P., & Sirothia, S. K. 2008, MNRAS, 383, 173

Kapferer, W., et al. 2007, A&A, 466, 813

Kapferer, W., Kronberger, T., Ferrari, C., Riser, T., & Schindler, S. 2008, MNRAS, 389, 1405

Kauffmann, G. 1995, MNRAS, 274, 153

Kauffmann, G., White, S. D. M., Heckman, T. M., Ménard, B., Brinchmann, J., Charlot, S., Tremonti, C., & Brinkmann, J. 2004, MNRAS, 353, 713

Kenney, J. D., & Young, J. S. 1986, ApJ, 301, L13

Kenney, J. D. P., Rubin, V. C., Planesas, P., & Young, J. S. 1995, ApJ, 438, 135

Kenney, J. D. P., Koopmann, R. A., Rubin, V. C., & Young, J. S. 1996, AJ, 111, 152

Kenney, J. D. P., van Gorkom, J. H., & Vollmer, B. 2004, AJ, 127, 3361

Kenney, J. D. P., Tal, T., Crowl, H. H., Feldmeier, J., & Jacoby, G. H. 2008, ApJ, 687, L69

Kennicutt, R. C., Jr. 1983, AJ, 88, 483

Kennicutt, R. C., Jr. 1998a, ARA&A, 36, 189

Kennicutt, R. C., Jr. 1998b, ApJ, 498, 541

Kern, K. M., Kilborn, V. A., Forbes, D. A., & Koribalski, B. 2008, MNRAS, 384, 305

Kilborn, V. A., Forbes, D. A., Barnes, D. G., Koribalski, B. S., Brough, S., & Kern, K. 2009, MNRAS, 400, 1962

Koopmann, R. A., & Kenney, J. D. P. 1998, ApJ, 497, L75

Koopmann, R. A., & Kenney, J. D. P. 2004a, ApJ, 613, 851

Koopmann, R. A., & Kenney, J. D. P. 2004b, ApJ, 613, 866

Kraan-Korteweg, R. C., Woudt, P. A., Cayatte, V., Fairall, A. P., Balkowski, C., & Henning, P. A. 1996, Nature, 379, 519

Kronberger, T., Kapferer, W., Ferrari, C., Unterguggenberger, S., & Schindler, S. 2008, A&A, 481, 337

Kulkarni, S. R., & Heiles, C. 1988, in Galactic and Extragalactic Radio Astronomy (A89-40409 17-90) (2nd ed.; Berlin and New York: Springer), 95

Landau, L. D., & Lifshitz, E. M. 1959, Fluid Mechanics (London: Pergamon), Chap. 9

Landau, L. D., & Lifshitz, E. M. 1960, Electrodynamics of Continuous Media (New York: Pergamon)

Larson, R. B., Tinsley, B. M., & Caldwell, C. N. 1980, ApJ, 237, 692

Lee, H., McCall, M. L., & Richer, M. G. 2003, AJ, 125, 2975

Leon, S., Combes, F., & Menon, T. K. 1998, A&A, 330, 37

Leroy, A. K., Walter, F., Brinks, E., Bigiel, F., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2782

Lewis, I., et al. 2002, MNRAS, 334, 673

Lisker, T., Grebel, E. K., Binggeli, B., & Glatt, K. 2007, ApJ, 660, 1186

Lisker, T., Grebel, E. K., & Binggeli, B. 2008, AJ, 135, 380

Machacek, M., Dosaj, A., Forman, W., Jones, C., Markevitch, M., Vikhlinin, A., Warmflash, A., & Kraft, R. 2005, ApJ, 621, 663

Machacek, M., Jones, C., Forman, W. R., & Nulsen, P. 2006, ApJ, 644, 155

Marcolini, A., Brighenti, F., & D'Ercole, A. 2003, MNRAS, 345, 1329

Margoniner, V. E., de Carvalho, R. R., Gal, R. R., & Djorgovski, S. G. 2001, ApJ, 548, L143

Mazure, A., Casoli, F., Durret, F., & Gerbal, D. 1998, in A New Vision of an Old Cluster: Untangling Coma Berenices. Proceedings, ed. A. Mazure, F. Casoli, F. Durret, & D. Gerbal (Singapore: World Scientific)

McGee, S. L., Balogh, M. L., Henderson, R. D. E., Wilman, D. J., Bower, R. G., Mulchaey, J. S., & Oemler, A., Jr. 2008, MNRAS, 387, 1605

McKee, C. F. 1995, in ASP Conference Series, Vol. 80, ed. A. Ferrara, C. F. McKee, C. Heiles, & P. R. Shapiro (San Francisco: Astronomical Society of the Pacific), 292

Mei, S., et al. 2007, ApJ, 655, 144

Mendes de Oliveira, C., & Hickson, P. 1994, ApJ, 427, 684

Mieske, S., Hilker, M., Infante, L., & Jordán, A. 2006, AJ, 131, 2442

Mieske, S., et al. 2008, A&A, 487, 921

Mihos, J. C., Harding, P., Feldmeier, J., & Morrison, H. 2005, ApJ, 631, L41

Minchin, R., et al. 2007, ApJ, 670, 1056

Moles, M., del Olmo, A., Perea, J., Masegosa, J., Marquez, I., & Costa, V. 1994, A&A, 285, 404

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613

Moore, B., Lake, G., Quinn, T., & Stadel, J. 1999, MNRAS, 304, 465

Moran, S. M., Ellis, R. S., Treu, T., Smith, G. P., Rich, R. M., & Smail, I. 2007, ApJ, 671, 1503

Mori, M., & Burkert, A. 2000, ApJ, 538, 559

Moss, C., Whittle, M., & Pesce, J. E. 1998, MNRAS, 300, 205

Mulchaey J. S., Zabludoff A. I., 1998, ApJ, 496, 73

Mulchaey, J. S., Davis, D. S., Mushotzky, R. F., & Burstein, D. 2003, ApJS, 145, 39

Murphy, E. J., Kenney, J. D. P., Helou, G., Chung, A., & Howell, J. H. 2009, ApJ, 694, 1435

Neistein, E., Maoz, D., Rix, H.-W., & Tonry, J. L. 1999, AJ, 117, 2666

Niklas, S. 1997, A&A, 322, 29

Niklas, S., Klein, U., & Wielebinski, R. 1997, A&A, 322, 19

Nulsen, P. E. J. 1982, MNRAS, 198, 1007

Oosterloo, T., & van Gorkom J. 2005, A&A, 437, L19

Osmond, J. P. F., & Ponman, T. J. 2004, MNRAS, 350, 1511

Panuzzo, P., et al. 2007, ApJ, 656, 206

Paolillo, M., Fabbiano, G., Peres, G., & Kim, D.-W. 2002, ApJ, 565, 883

Pappalardo, C., Lançon, A., Vollmer, B., Ocvirk, P., Boissier, S., & Boselli, A. 2010, A&A, 514, A33

Paudel, S., Lisker, T., Kuntschner, H., Grebel, E. K., & Glatt, K. 2010, MNRAS, 405, 800

Phillipps, S., Drinkwater, M. J., Gregg, M. D., & Jones, J. B. 2001, ApJ, 560, 201

Phookun, B., Vogel, S. N., & Mundy, L. G. 1993, ApJ, 418, 113

Pilyugin, L. S., Mollá, M., Ferrini, F., & Vílchez, J. M. 2002, A&A, 383, 14

Poggianti, B. M., Smail, I., Dressler, A., Couch, W. J., Barger, A. J., Butcher, H., Ellis, R. S., & Oemler, A., Jr. 1999, ApJ, 518, 576

Poggianti, B. M., Bridges, T. J., Komiyama, Y., Yagi, M., Carter, D., Mobasher, B., Okamura, S., & Kashikawa, N. 2004, ApJ, 601, 197

Poggianti, B. M., et al. 2006, ApJ, 642, 188

Popescu, C. C., Tuffs, R. J., Völk, H. J., Pierini, D., & Madore, B. F. 2002, ApJ, 567, 221

Postman, M., et al. 2005, ApJ, 623, 721

Quilis, V., Moore, B., & Bower, R. 2000, Science, 288, 1617

Randall, S., Nulsen, P., Forman, W. R., Jones, C., Machacek, M., Murray, S. S., & Maughan, B. 2008, ApJ, 688, 208

Rasmussen, J., Ponman, T. J., Verdes-Montenegro, L., Yun, M. S., & Borthakur, S. 2008, MNRAS, 388, 1245

Rasmussen, J., Sommer-Larsen, J., Pedersen, K., Toft, S., Benson, A., Bower, R. G., & Grove, L. F. 2009, ApJ, 697, 79

Rengarajan, T. N., Karnik, A. D., & Iyengar, K. V. K. 1997, MNRAS, 290, 1

Roediger, E., & Brüggen, M. 2006, MNRAS, 369, 567

Roediger, E., & Brüggen, M. 2007, MNRAS, 380, 1399

Roediger, E., & Brüggen, M. 2008, MNRAS, 388, 465

Roediger, E., Brüggen, M., & Hoeft, M. 2006, MNRAS, 371, 609

Roediger, E., & Hensler, G. 2005, A&A, 433, 875

Roediger, E., & Hensler, G. 2008, A&A, 483, 121

Sabatini, S., Davies, J., van Driel, W., Baes, M., Roberts, S., Smith, R., Linder, S., & O'Neil, K. 2005, MNRAS, 357, 819

Salpeter, E. E., & Hoffman, G. L. 1996, ApJ, 465, 595

Sancisi, R., Fraternali, F., Oosterloo, T., & van der Hulst, T. 2008, A&ARv, 15, 189

Sarazin, C. L. 1986, RvMP, 58, 1

Schindler, S., Binggeli, B., & Böhringer, H. 1999, A&A, 343, 420

Schmidt, M. 1963, ApJ, 137, 758

Schoenmakers, R. H. M., Franx, M., & de Zeeuw, P. T. 1997, MNRAS, 292, 349

Schulz, S., & Struck, C. 2001, MNRAS, 328, 185

Sengupta, C., Balasubramanyam, R., & Dwarakanath, K. S. 2007, MNRAS, 378, 137

Sivanandam, S., Rieke, M. J., & Rieke, G. H. 2010, ApJ, 717, 147

Skillman, E. D., Kennicutt, R. C., Jr., Shields, G. A., & Zaritsky, D. 1996, ApJ, 462, 147

Smith, G. P., Treu, T., Ellis, R. S., Moran, S. M., & Dressler, A. 2005, ApJ, 620, 78

Smith, R. J., et al. 2010, MNRAS, 408, 1417

Soida, M., Otmianowska-Mazur, K., Chyży, K., & Vollmer, B. 2006, A&A, 458, 727

Solanes, J. M., Manrique, A., García-Gómez, C., González-Casado, G., Giovanelli, R., & Haynes, M. P. 2001, ApJ, 548, 97

Spitzer, L., Jr. 1978, Physical Processes in the Interstellar Medium (New York: Wiley)

Spitzer, L., Jr. 1990, ARA&A, 28, 71

Stark, A. A., Knapp, G. R., Bally, J., Wilson, R. W., Penzias, A. A., & Rowe, H. E. 1986, ApJ, 310, 660

Stevens, I. R., Acreman, D. M., & Ponman, T. J. 1999, MNRAS, 310, 663

Stickel, M., Bregman, J. N., Fabian, A. C., White, D. A., & Elmegreen, D. M. 2003, A&A, 397, 503

Struck, C. 1999, PhR, 321, 1

Sulentic, J. W., Rosado, M., Dultzin-Hacyan, D., Verdes-Montenegro, L., Trinchieri, G., Xu, C., & Pietsch, W. 2001, AJ, 122, 2993

Sun, M., & Vikhlinin, A. 2005, ApJ, 621, 718

Sun, M., Jones, C., Forman, W., Nulsen, P. E. J., Donahue, M., & Voit, G. M. 2006, ApJ, 637, L81

Sun, M., Donahue, M., & Voit, G. M. 2007, ApJ, 671, 190

Sun, M., Donahue, M., Roediger, E., Nulsen, P. E. J., Voit, G. M., Sarazin, C., Forman, W., & Jones, C. 2010, ApJ, 708, 946

Takeda, H., Nulsen, P. E. J., & Fabian, A. C. 1984, MNRAS, 208, 261

Tanaka, M., Goto, T., Okamura, S., Shimasaku, K., & Brinkmann, J. 2004, AJ, 128, 2677

Thomas, T., & Katgert, P. 2006, A&A, 446, 31

Tielens, A. G. G. M., & Hollenbach, D. 1985, ApJ, 291, 722

Toloba, E., Boselli, A., Cenarro, A. J., Peletier, R. F., Gorgas, J., Gil de Paz, A., & Muñoz-Mateos, J. C. 2011, A&A, 526, A114

Tonnesen, S., Bryan, G. L., & van Gorkom, J. H. 2007, ApJ, 671, 1434

Tonnesen, S., & Bryan, G. L. 2009, ApJ, 694, 789

Tonnesen, S., & Bryan, G. L. 2010, ApJ, 709, 1203

Tonnesen, S., Bryan, G. L., & Chen, R. 2011, ApJ, 731, 98

Toomre, A., & Toomre, J. 1972, ApJ, 178, 623

Tosa, M. 1994, ApJ, 426, L81

Treu, T., Ellis, R. S., Kneib, J.-P., Dressler, A., Smail, I., Czoske, O., Oemler, A., & Natarajan, P. 2003, ApJ, 591, 53

Urquhart, S. A., Willis, J. P., Hoekstra, H., & Pierre, M. 2010, MNRAS, 406, 368

Valluri, M. 1993, ApJ, 408, 57

van den Bergh, S. 2009, ApJ, 702, 1502

van Zee, L., Skillman, E. D., & Haynes, M. P. 2004a, AJ, 128, 121

van Zee, L., Barton, E. J., & Skillman, E. D. 2004b, AJ, 128, 2797

Verdes-Montenegro, L., Yun, M. S., Perea, J., del Olmo, A., & Ho, P. T. P. 1998, ApJ, 497, 89

Verdes-Montenegro, L., Yun, M. S., Williams, B. A., Huchtmeier, W. K., Del Olmo, A., & Perea, J. 2001, A&A, 377, 812

Verheijen, M., van Gorkom, J. H., Szomoru, A., Dwarakanath, K. S., Poggianti, B. M., & Schiminovich, D. 2007, ApJ, 668, L9

Vikhlinin, A., Markevitch, M., & Murray, S. S. 2001, ApJ, 549, L47

Völk, H. J., & Xu, C. 1994, InPhT, 35, 527

Vollmer, B. 2003, A&A, 398, 525

Vollmer, B. 2009, A&A, 502, 427

Vollmer, B., & Beckert, T. 2002, A&A, 382, 872

Vollmer, B., Cayatte, V., Balkowski, C., & Duschl, W. J. 2001a, ApJ, 561, 708

Vollmer, B., Braine, J., Balkowski, C., Cayatte, V., & Duschl, W. J. 2001b, A&A, 374, 824

Vollmer, B., & Huchtmeier, W. 2003, A&A, 406, 427

Vollmer, B., Thierbach, M., & Wielebinski, R. 2004a, A&A, 418, 1

Vollmer, B., Beck, R., Kenney, J. D. P., & van Gorkom, J. H. 2004b, AJ, 127, 3375

Vollmer, B., Huchtmeier, W., & van Driel, W. 2005a, A&A, 439, 921

Vollmer, B., Braine, J., Combes, F., & Sofue, Y. 2005b, A&A, 441, 473

Vollmer, B., Soida, M., Otmianowska-Mazur, K., Kenney, J. D. P., van Gorkom, J. H., & Beck, R. 2006, A&A, 453, 883

Vollmer, B., Soida, M., Beck, R., Urbanik, M., Chyży, K. T., Otmianoska-Mazur, K., Kenney, J. D. P., & van Gorkom, J. H. 2007, A&A, 464, L37

Vollmer, B., Braine, J., Pappalardo, C., & Hily-Blant, P. 2008, A&A, 491, 455

Vollmer, B., Soida, M., Chung, A., Chemin, L., Braine, J., Boselli, A., & Beck, R. 2009, A&A, 496, 669

Vollmer, B., Soida, M., Chung, A., Beck, R., Urbanik, M., Chyży, K. T., Otmianowska-Mazur, K., & van Gorkom, J. H., 210, A&A, 512, A36

Vollmer, B., et al. 2012, A&A, 537, A143

Walter, F., Brinks, E., de Blok, W. J. G., Bigiel, F., Kennicutt, R. C., Jr., Thornley, M. D., & Leroy, A. 2008, AJ, 136, 2563

Warmels, R. H. 1988, A&AS, 73, 453

Whitmore, B. C., & Gilmore, D. M. 1991, ApJ, 367, 64

Whitmore, B. C., Gilmore, D. M., & Jones, C. 1993, ApJ, 407, 489

Williams, B. A., Yun, M. S., & Verdes-Montenegro, L. 2002, AJ, 123, 2417

Wilman, D. J., et al. 2005, MNRAS, 358, 88

Woudt, P. A. 1998, Ph.D. thesis, University of Cape Town, South Africa

Woudt, P. A., Kraan-Korteweg, R. C., Lucey, J., Fairall, A. P., & Moore, S. A. W. 2008, MNRAS, 383, 445

Yagi, M., et al. 2010, AJ, 140, 1814

Yoshida, M., et al. 2004, AJ, 127, 90

Yoshida, M., et al. 2008, ApJ, 688, 918

Zabludoff, A. I., & Mulchaey, J. S. 1998, ApJ, 496, 39

Zepf, S. E., & Whitmore, B. C. 1991, ApJ, 383, 542

6 Clusters of Galaxies

Richard Bower
Department of Physics, University of Durham, Durham, UK

1 Introduction *267*
1.1 What Is a Cluster? 267
1.2 Historical Perspective 267
1.3 Overview 271

2 The Optical Properties of Clusters *271*
2.1 The Density–Morphology Relation 271
2.2 The Color–Magnitude Relation 273
2.3 Spectroscopic Properties of Cluster Galaxies 274
2.4 The Fundamental Plane of Early-Type Galaxies 275
2.5 Galaxy Ecology 276
2.5.1 Galaxy Collisions 276
2.5.2 Dynamical Friction 277
2.5.3 Ram-Pressure Stripping 277
2.5.4 Strangulation 279
2.6 Evolution of Galaxy Clusters 279
2.6.1 Color Evolution 279
2.6.2 E+A Galaxies 281
2.6.3 Cluster Archaeology 281
2.6.4 The Luminosity Function 281
2.6.5 Morphology 283
2.6.6 Other Wavebands 284
2.7 The Relation of Clusters to Galaxy Groups 284
2.8 Intra-cluster Light 285

3 X-Ray Emission *285*
3.1 The Physics of X-Ray Emission 285
3.1.1 Thermal Bremsstrahlung Emission 285
3.1.2 Bound–Bound Electron Transitions 286
3.1.3 Total Emissivity 288
3.2 The Baryon Content of Galaxy Clusters 288
3.2.1 Temperature and Density Profiles 288
3.2.2 The Entropy Distribution 289
3.3 The Cooling Instability 291
3.3.1 Cooling Flows: Comparison to Observations 293
3.3.2 Resolution 293
3.4 The Sunyaev–Zeldovich Effect 295

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_6, © Springer Science+Business Media Dordrecht 2013

4 Dark Matter *295*

4.1 Galaxy Dynamics 295

4.2 Hydrostatic Equilibrium 296

4.3 Gravitational Lensing 297

4.4 Implications for Cosmology 298

5 The Formation of Clusters *299*

6 Summary and the Future *301*

6.1 Summary 301

References *302*

Abstract: This chapter focuses on galaxy clusters, tackling three aspects. Firstly, we look at clusters as laboratories, where we can study galaxy evolution and the interaction of galaxies with their environment. By measuring the properties of galaxies in clusters as a function of redshift, we can build a statistical history of galaxy formation and test this against theoretical models. Secondly, we look at the diffuse intra-cluster medium (ICM) that pervades galaxy clusters. This plasma makes accounts for the vast majority of baryons in the cluster. X-ray emission from the plasma can be used to measure the gravitational potential and to trace the thermal history of the universe. Finally, we look at clusters as probes of cosmology and their role in complementing the cosmic microwave background in constraining cosmological parameters.

1 Introduction

1.1 What Is a Cluster?

What is a cluster? Clusters are self-gravitating systems of galaxies, hot gas, and dark matter. By convention, they have typical masses of 10^{14}–$10^{15}\,M_\odot$ and contain 100s–1,000s of galaxies. We distinguish clusters by their mass. They are the most massive collapsed objects in the present-day universe. Clusters are rare, and even the least massive have a space density less than 10^{-5} Mpc^{-3}. The most massive examples have a space density of ~10^{-8} Mpc^{-3} in the local universe. The space density declines rapidly to higher redshifts. Our nearest rich cluster, the Coma cluster, is a good example of fairly typical galaxy cluster (see ➋ *Fig. 6-1*). Its close proximity makes it possible to study the galaxy properties in great detail.

We often refer to lower mass (and more common) systems with masses in the range $10^{14}\,M_\odot > M_{\rm tot} > 10^{13}\,M_\odot$ as galaxy groups. There is no strict definition of the mass limits, however, and some authors would include the systems containing two bright galaxies (e.g., the Milky Way and Andromeda) as "groups." Many papers also refer to "field" galaxies. Such galaxies are really average galaxies drawn from random locations in the universe. Consequently, some field galaxies might well lie in groups, while clusters are too rare to be represented. It is perhaps better to refer to "isolated" galaxies if we want to contrast the properties of galaxies inside and outside groups.

Our modern understanding is that clusters form as the result of gravitational instability. They are the result of small density fluctuations in the post-inflation universe. The gravitational instability creates a whole spectrum of virialized (i.e., stable) dark matter haloes. As the universe evolves, the characteristic halo mass moves to larger and larger scales. Clusters are the most massive of these haloes and exponentially rare in the present-day universe. They will become increasingly common in the future, until the acceleration driven by dark energy eventually halts development of cosmic structure.

1.2 Historical Perspective

It is helpful to begin with a brief review of the long history of galaxy cluster studies. Almost as soon as the first astronomical telescopes were developed, clusters were identified as striking concentrations of galaxies on the sky. The discovery predated the realization that the "spiral

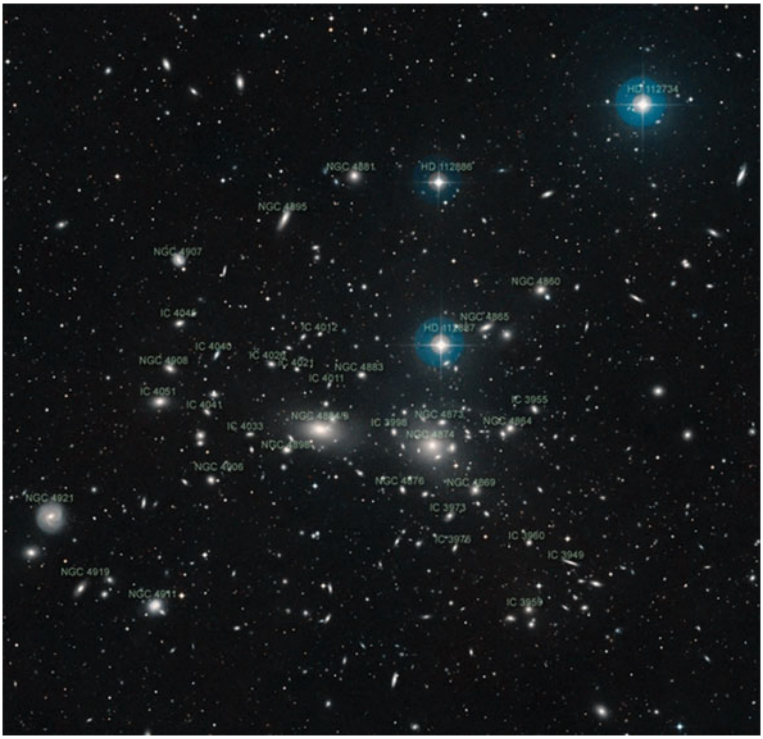


Fig. 6-1
With a mass of 2 × 10^{15} $M_{\odot}$, the Coma cluster is our nearest large cluster. It lies at a distance of 100 Mpc. The image shows the central regions of the cluster and is annotated with some of the brighter galaxies catalogued by early astronomers. The cluster is often used as the archetypal cluster. Nevertheless, the cluster may have resulted from a recent merger of two smaller systems resulting in the two dominant galaxies, NGC 4884 and NGC 4874 (Image credit: NASA/DXS)

nebulae" were outside our own galaxies. A clear division in galaxy properties was already evident in these observations, and it was immediately evident that most galaxies in clusters were amorphous systems, devoid of spiral arms. This density–morphology relation has been the motivation of a great deal of "ecological" research over past decades. The central questions have morphed into a consideration of the star formation histories of galaxies, but the issues are still the same.

The development of spectrographs allowed the redshifts (and hence relative velocities) of galaxies to be measured. This lead to an apparent contradiction: if clusters were bound, the mass associated with each of the galaxies must greatly exceed the visible stellar mass. Zwicky had discovered the existence of dark matter. Of course, these observations do not make it clear that the dark matter is non-baryonic. The discovery of the hot gas in clusters had to wait until the development of X-ray balloon observations. The discovery of bright X-ray emission was a great surprise, but it was quickly realized that the mass of hot gas trapped by the gravitational potential greatly exceeded that in visible stars.

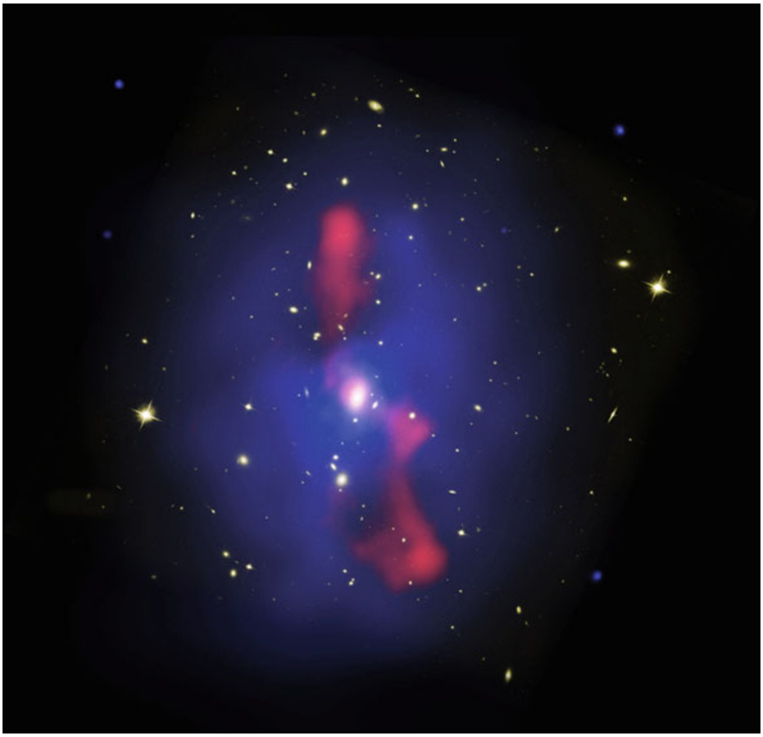
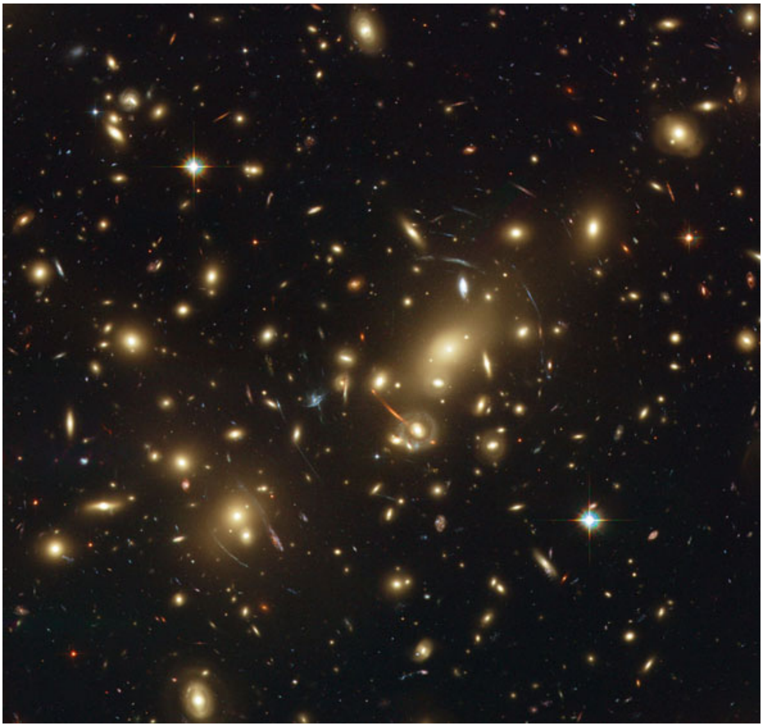


Fig. 6-2
An example of the spectacular interaction between a radio galaxy and the surrounding cluster. This image shows the cluster MS0735.6 + 7421 (McNamara et al. 2005) and superposes the optical (*yellow*), X-ray (*blue*), and radio (emission). The jet of radio emission from the central galaxy has swept aside the intra-cluster medium creating a cavity in the X-ray-emitting plasma. Such events are hugely energetic and are thought to balance the radiative cooling of the intra-cluster medium (Image credit: NASA/CXC/SAO)

The discovery of hot gas, however, leads to a new puzzle. The luminosity implied a high cooling rate. Surely, this cooling gas must go somewhere, and surely, it must lead to the formation of stars. The "cooling flow" paradox was born. A paradox because the cooling rates seemed incompatible with the red colors and low rates of star formation in cluster galaxies. Some authors suggested that the gas might form rocks or invisible low-mass stars, but resolution would have to wait for the vastly improved image quality and spectroscopic energy resolution of X-ray satellites.

A key piece of the puzzle is the frequent association of "radio galaxies" with clusters. Such galaxies have strong radio emission, frequently in the form of jets generated by accretion onto a central black hole. An example is shown in *Fig. 6-2*. Thus, the story of galaxy clusters links the collapse of dark matter on cosmological (100 Mpc or 10^{24} m) scales to the accretion physics on the scale of the black hole horizon (scales of 10^{11} m).

More recently, gravitational lensing has provided overwhelming evidence of the existence of dark matter. Although the bending of light was an early prediction of general relativity, it was



◘ **Fig. 6-3**
The galaxy cluster A2218. The image highlights the spectacular gravitational lensing created by gravitational potential of this cluster. The arc-like features in this image are the distorted images of background galaxies. The source is strongly magnified making, it possible to exploit the cluster as a gravitational telescope. In addition to this strong lensing effect, the cluster weakly distorts the shapes of background objects across the entire field. Inverting this distortion provides a means of directly measuring the cluster's mass (Image credit: NASA/HST)

not widely appreciated that clusters were so massive that they would easily bend light from background objects into spectacular giant arcs. Initial explanations suggested that these were galaxies being ripped apart by the tidal forces of the cluster, but spectroscopy clearly showed them to be objects at much greater distance than the cluster. Spectacular arcs have now been observed from young galaxies at $z = 5$ and greater (➋ *Fig. 6-3*).

The discovery of gravitational lensing has ushered in a new era of cluster studies, where the mass content and concentration of galaxy clusters can be determined without reference to the galaxy content or the X-ray hot gas. This has made it possible to directly calibrate the masses of clusters and thus to use them as probes of galaxy evolution. More startlingly, the clusters can also become telescopes, gravity creating the first lens of a cosmic telescope that provides an unprecedented views of the first galaxies to form in the universe.

1.3 Overview

In this chapter, I will take the reader on a guided tour of forefront studies of galaxy clusters. This will begin with a close look at the optical properties of clusters and the galaxies they contain. I will examine the special morphologies and star formation histories of cluster galaxies compared to those in more isolated environments. The aim is to understand how the differences between these galaxies are established. This has led to better understanding of Galaxy Ecology – the ways in which galaxies interact with their surroundings – and to improved understanding of the suppression of star formation in the universe in general. An important topic at the present is to compare the properties of galaxies in presented day clusters with those of clusters at higher redshifts. The higher redshift clusters are seen at earlier times and are (in a statistical sense) the progenitors of toady's most massive systems. This is a fertile test-bed for theories of galaxy formation and evolution.

Although clusters are often thought of as collections of galaxies, galaxies are far from the dominant baryonic component. A far greater fraction of the system's baryons are associated with a diffuse hot plasma that is confined by the gravitational pull of the system's dark matter. The plasma is a strong X-ray emitter, and the spectrum can be analyzed to reveal the plasma density and temperature. This allows the cluster to be used as cosmic calorimeter reflecting the thermal history of its formation. I will discuss the major puzzle of this plasma – the "cooling flow paradox." Although the plasma has a short cooling time, the measured star formation rates of cluster galaxies are much smaller than this would seem to imply. The resolution to this paradox seems to be intimately connected to the frequent presence of radio galaxies in cluster of galaxies, an observation that has now been incorporated into theoretical models of galaxy formation with profound consequences.

In the final section, I will look at clusters of galaxies from a theoretical perspective, focusing on the evidence for their dark matter content and on their role in setting cosmological parameters. I will also set the formation of clusters of galaxies in the context of the growth of the large-scale structure of the cosmos.

2 The Optical Properties of Clusters

2.1 The Density–Morphology Relation

Galaxies contain two main components – a flattened stellar disk that is supported by the coherent rotation of stars and an ellipsoidal bulge that is supported the random and roughly isotropic motion of its stars. Hubble developed a system for classifying galaxies on the basis of (1) the relative strength of the disk and bulge components and (2) the presence and strength of spiral arms. Galaxies with a dominant bulge component and weak spiral arms are referred to as "early type" while the galaxies with a strong disk and clear spiral arms are "late types." The designation is not intended to indicate a morphological sequence. The late or spiral galaxies are then subclassified according to the presence of a central bar as well as the strength of the bulge/spiral arms. An important class of galaxies is the lenticular or S0 galaxies. These have a clear disk component (as well as a strong bulge) but weak or absent spiral arms. Elliptical galaxies have significantly weaker (or undetectable) disks. From the view point of clusters, the key distinction is between the early and late types and between the elliptical and S0 galaxies.

Dressler (1980) applied the classification scheme to clusters of galaxies. This was a heroic work obtaining clear photographic plates and laboriously classifying the galaxies in the clusters. Although it had been clear to even the first observers (such as Herschel and Wolfe) that early types were clustered into the densest regions, Dressler's measurements made it possible to quantify this trend. The results showed a universal relationship between the local density and galaxy type, and Dressler used this to argue that there was a causal connection between galaxy morphology and clustering. More recent results have been used to show that the trends extend to lower density environments such as galaxy groups and that the development of the relation can be traced back to high redshift.

The origin of galaxy morphology is often phrased in terms of "nature vs. nurture." In other words, is the morphology of galaxies a result of the initial conditions of galaxy formation, or as the result of the later evolution of the galaxy? Were the galaxies formed differently and then captured into the cluster, or were they modified by the cluster? Current theories suggest that both play a role. The galaxies in clusters (and groups) tend to be more massive than galaxies in lower density environments. This describes a large part of the trend since more massive galaxies are very often elliptical. But a trend with environment is still evident at a fixed stellar mass, and this is ascribed to various physical processes that tend to suppress star formation in cluster galaxies (converting spirals to S0 types as the disk fades and the arms dissolve) and to randomize the coherent disk motion to form a bulge (converting S0 to elliptical types). Ram-pressure stripping of disk and halo gas is thought to drive the spiral to S0 transformation, while the S0 to E transformation is most likely driven by galaxy collisions/mergers and harassment (the cumulative effect of weak encounters). Theoretical models for galaxy evolution make an important distinction between the galaxy at the center of the system and the satellite galaxies that orbit around it. We should expect the satellites to be subject to the transformation processes, while the central galaxy may continue to accrete further gas from its surroundings. I will describe the transformation processes in more detail below. The timescales for the two transitions need not be the same.

At higher redshift, the trends evolve, with higher redshift clusters containing a higher proportion of later-type galaxies. A particularly contentious issue has been the rapid buildup of the S0 content of clusters. The controversy has largely arisen due to the difficulty of clearly distinguishing the E and S0 types and in allowing for the biases in clusters/galaxy selection due to the redshifting of the observed bands.

A major effort has therefore been made to establish a morphological classification scheme that can be determined by computers rather than the human eye. Machine-based schemes use the light profile or the concentration and asymmetry (Simard et al. 2002). Some success has been obtained in describing the broad-brush evolution of the galaxy population as a whole, but the differences in the morphologies of cluster galaxies are rather more subtle. Indeed, the issue of mapping machine-based classifications onto Hubble scheme is difficult because the original classification scheme relies on more than one property.

A new development is the "Galaxy Zoo" (Lintott et al. 2008; Bamford et al. 2009): using a popular web site, "citizen-scientists" are invited to classify galaxy images (currently from the Sloan Digital Sky Survey). Although each individual classifier may have a large uncertainty in classifying one object, each galaxy is classified by hundreds of people so that the trends can be carefully examined with statistical tools. Recent results present a reassuring confirmation of the trends identified by Dressler but emphasize that galaxy mass has at least as important a role as galaxy environment.

2.2 The Color–Magnitude Relation

The lack of a clear physical interpretation of galaxy morphology has led many authors to prefer classification schemes based on galaxies, stellar populations and star formation histories. In outline, spiral galaxies have significant contributions from young, relatively hot stars (this makes the spiral arms stand out in imaging data), while elliptical and S0 galaxies are dominated by older stars similar to the sun. Color (or certain spectral absorption lines) may therefore be used to create a classification scheme that is clearly related to underlying physical properties of the galaxies. The universe encourages us to use this scheme since galaxies fall onto two largely separated sequences of passive galaxies (with red colors and little ongoing star-formation) and star-forming galaxies (with blue colors and specific star formation rates around 0.1 Gyr^{-1}, such galaxies will double their stellar mass in 10 Gyr which suggests that their average star formation rate has changed little over the history of the universe).

Galaxy color has the advantage that it is relatively "cheap" (in terms of telescope exposure time) to measure and that the average color of the whole galaxy can be determined (making it simple to compare galaxies of different angular size). Colors have the drawback that they are affected by both the age of the stellar population and its star formation history. This degeneracy can make it hard to interpret results uniquely, although the situation is improved by using several colors spanning along wavelength base line, preferably including data from the near-infrared. It is also difficult to determine instantaneous star formation rates from optical/near-IR data: so, additional data in the ultraviolet or mid-infrared is required. Spectra of galaxies contain more information than galaxy colors, allowing instantaneous star formation rates to be determined from strong nebular emission lines (such as Hα) and allowing the age-metalicity degeneracy to be broken using carefully selected stellar absorption features. This is an extensive topic. Nevertheless, a great deal can be learned from a look at the broadband optical colors of cluster galaxies, particularly if the galaxy redshift is known (or can be accurately estimated using multiple photometric bands).

The standard approach is to plot galaxy color as a function of luminosity. In clusters, a clear sequence of red galaxies stands out. An example (for the Coma cluster) is shown in ➋ *Fig. 6-4*. All the galaxies in this plot are confirmed to have redshifts matching that of the cluster. The close proximity of the cluster makes it possible to measure accurate colors for even very faint galaxies and to show that the sequence extends over a range of 10,000 in stellar mass.

The existence of this relation places strong constraints on the formation history of cluster galaxies, as well as making clusters stand out in imaging of random fields. The relation is so red because the majority of the cluster galaxies have few young stars: the bulk of the stars in these systems must have formed many Gyr in the past. If we focus on the cores of galaxy clusters, the very small proportion of cluster galaxies that deviate to the blue side of this sequence implies that few star-forming galaxies are present and the fraction of blue galaxies can be used to establish an analog of the morphology–density relation. The slope of the red sequence can be understood in terms of a stellar mass – metalicity and stellar mass – age relationships. These relationships appear to be universal from cluster to cluster, implying a great deal of similarity in cluster formation history.

The cluster can be contrasted with a random field where the galaxy colors are much more smoothly distributed in the color plane. Although a red-sequence is also evident in the random field, it is occupied by a far smaller fraction of galaxies (recall that the random field will include galaxy groups in the sample). However, the spread in the red shifts of the field systems means

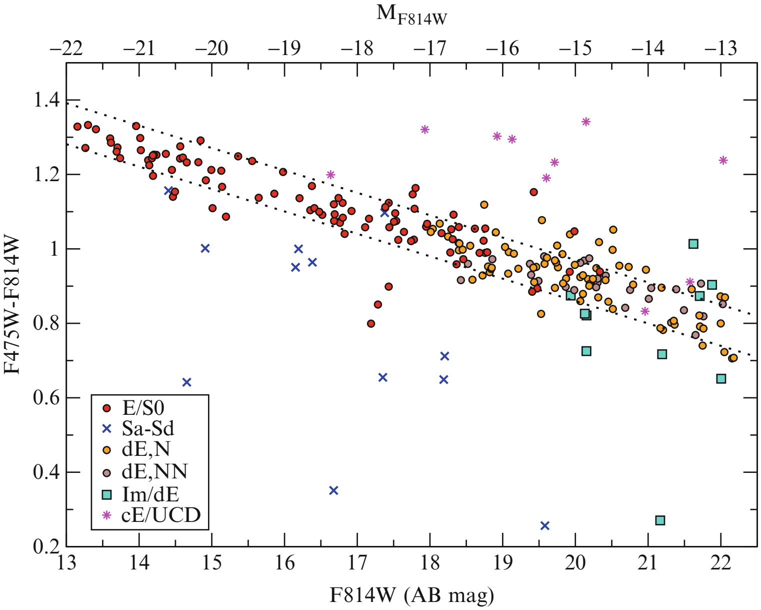


Fig. 6-4

The color-magnitude relation in the Coma cluster. The points show the strong correlation between the brightness of cluster galaxies and their color. The points are labeled by morphology. The relation obeyed by the early-type (E/S0 and dwarf E) galaxies is extremely tight with a scatter of less than 0.06 mag. In these spectacular observations, the relation is seen to hold over 10 mag in galaxy luminosity (a factor of 10,000 in galaxy mass) (Image credit: Hammer et al. 2010)

that the sequence (and the corresponding bluer sequence of star-forming galaxies) is smeared out. This makes it possible to effectively select clusters of galaxies using optical techniques, capitalizing on the enhanced the contrast of the cluster against the projected background of the field galaxies (Gladders and Yee 2000).

2.3 Spectroscopic Properties of Cluster Galaxies

The spectra of galaxies are, of course, essential for confirming a galaxy's membership of a cluster, but they also contain a great deal of information about the current star formation rate of galaxies, their past star formation histories, and their metal abundance.

- *Star formation rate diagnostics*. Ongoing star formation rates of galaxies can be estimated from the emission lines in the spectrum. The technique measures the total ionizing flux produced by massive stars, since this is remitted as line radiation when the interstellar gas recombines. Because hydrogen is so abundant (and has a low-ionization potential), it is particularly strong. One of the best star formation tracers is thus the Hα emission line at 6,563 Å (Kennicutt 1998; Brinchmann et al. 2004). Conversion of a line luminosity to a star formation rate is, however, complicated by the presence of dust that may scatter or absorb line photons before they leave the galaxy. It is possible to correct for this using the relative strength of Hα and Hβ. Other elements, notably oxygen, also produce strong lines, but interpretation of these lines as star formation indicators is dependent on the system's metal abundance (as well as dust), so they must be carefully calibrated. It is also possible to estimate star formation rates on the basis of the mid-IR (e.g., 24 μ) flux. This uses the flux of UV radiation that is reprocessed by dust. In this way, it is possible to determine star formation rates even in galaxies that show no detectable emission. With all these methods, care needs to be taken that the source of flux is star formation rather than accretion on to a black hole (an AGN).

- *Star formation history diagnostics*. Absorption lines in the spectrum can be used to study the star formation history of a galaxy. This is usually parameterized as the luminosity weighted age of the stellar population. The technique relies on the different surface gravity and metalicity sensitivity of various Balmer and metal lines (Worthey 1994). Strong higher-order Balmer lines are indicative of a young stellar population. For example, an Hδ equivalent width of greater than 3 Å can only be explained by the presence of a substantial population of A-class stars. This indicates a luminosity weighted stellar age of around 1 Gyr.
- *Metalicity indicators*. The metal abundance of the stellar population can be determined from diagnostic lines in a similar way. However, the differences between lines as a function of metalicity are relatively subtle and careful modeling is required. In particular, galaxies in clusters often show an enhancement in the abundance of α process elements (such as Mg) relative to that in the solar spectrum. This provides an additional measure of the star formation history of the galaxy suggesting a relatively rapid enrichment, primarily by type II supernovae.

These techniques allow the galaxy population to be described in more detail. Current results confirm that the most massive early-type galaxies are old (age ~10 Gyr) and metal rich. There is a consistent tendency for lower luminosity galaxies to be both less metal rich and to be younger than the more massive galaxies. This may agree well with the results from the observed evolution of galaxy color–magnitude diagrams. In addition, the lower mass galaxies show less enrichment of α process elements, suggesting that they have a more extended star formation history as well as being younger (Nelan et al. 2005).

2.4 The Fundamental Plane of Early-Type Galaxies

The spectra of galaxies also allow the velocity dispersion or rotation speed of the stellar system to be determined. Combining this with the stellar radius of the system yields the dynamical mass of the galaxy. In clusters, this is most often applied to the elliptical and bulge-dominated S0 population. Since the velocity dispersion profile is approximately flat, the mass can be estimated by the simple application of the virial theorem. If we further assume that the mass is dominated by the stellar population (or at least proportional to the luminosity) and that the shape of the system is homologous, this predicts a tight correlation between the system radius, velocity dispersion, and mass:

$$R_e \propto \frac{M_e}{\sigma^2} \qquad \text{or} \qquad R_e \propto \sigma^2 I_e^{-1} \tag{6.1}$$

where R_e is the effective radius of the galaxy, M_e the mass within R_e, σ the central velocity dispersion, and I_e the surface brightness at R_e. Such a tight correlation is indeed observed, although the coefficients of the observed relation differ slightly (Jorgensen et al. 1996):

$$R_e \propto \sigma^{1.2} I_e^{-0.8} \tag{6.2}$$

The difference between the two appears to arise from variations in the mass-to-light ratio of the stellar population and a degree of nonhomology. Such variations are not surprising since the stellar populations are known to vary with total system mass. Indeed, many galaxy properties correlate more tightly with velocity dispersion than with system mass.

By choosing suitable combinations of the surface brightness and the velocity dispersion, the fundamental plane may be viewed edge on. The resulting tight correlation with luminosity may be used as a powerful distance indicator and also as a strong constraint on the mass-to-light ratio evolution of cluster galaxies (Holden et al. 2005).

2.5 Galaxy Ecology

Galaxy ecology is the study of how galaxies interact with their environment. Clusters make a good laboratory (Gunn and Gott 1972). There are lots of galaxies in one place, at one redshift. The drawback is that cluster galaxies represent only a small fraction of the total galaxy population. Galaxy groups contain a much larger fraction of the population, and galaxy transformations, also more likely there. The subject is opening up due to large redshift surveys which identify large samples of group galaxies, but at high redshift, these surveys are still limited to relatively bright galaxies.

Although the trends of morphology and color with environment density are well established, the theoretical underpinnings of the relationships are only now becoming clear. There are essentially four key processes that transform galaxy properties: galaxy collisions, dynamical friction, the ram pressure of the intra-cluster medium, and "strangulation." However, the properties of galaxies in clusters are not simply the result of galaxy transformation. Cluster galaxies are generally brighter than spirals, so the comparison needs care to ensure that galaxies are being compared on a like-for-like basis. Samples must match in stellar masses, not luminosities, and the comparison must allow for tidal stripping and fading of the disk. Finally, it should be remembered that the progenitors of cluster galaxies are not the spiral galaxies we observe today, but the field galaxy population at $z \sim 1$.

2.5.1 Galaxy Collisions

Consider what happens when one galaxy passes another. The stars feel a perturbation in their gravitational force. A rapid fluctuation removes energy from the (ordered) motion of the galaxy and puts it into random motions of the stars. If the encounter is sufficiently rapid, we can use impulsive approximation. The energy transferred from the orbital motion of the galaxies to the internal motion of the stars is then

$$\Delta E \sim \frac{G^2 M_2^2}{V_p^2 R_p^2} M_1 \left(\frac{r_1}{R_p} \right)^2 \tag{6.3}$$

where M_1 and M_2 are the masses of the perturbed and perturbing galaxies, respectively, r_1 is the size of the perturbed galaxy, their relative velocity is V_p, and their closest approach distance is R_p (Richstone 1975; Covington et al. 2008).

The effect is (1) to puff up the galaxy's stellar disk. If the energy transfer is extreme, the heating destroys the disk. S0 disks are indeed thicker than those of spiral galaxies. (2) Remove orbital angular momentum. If enough is removed, the two galaxies become bound, and as they encounter each other again, more energy is lost and they spiral together eventually merging to form an (elliptical) remnant. If the galaxy masses are comparable, the remnant will be dominated by the bulge. The remnant's shape is dominated by velocity dispersion (which need not be isotropic). If the masses are unequal, a significant disk may survive.

However, these processes are more effective if the encounter speed is comparable to the motion of the stars. If it is too rapid, the energy exchange is minimal. As a result, clusters with high-velocity dispersions are not likely to be strongly influenced by collisions. "Harassment," the cumulative effect on multiple, weak encounters, may be important however (Moore et al. 1996).

2.5.2 Dynamical Friction

As a galaxy travels through the dark matter halo of the cluster, it experiences a drag force. This results because the dark matter is focused towards the galaxy as it orbits and generates a wake behind the galaxy. The slight mass excess exerts a retarding force on the galaxy's motion.

Over time, this saps the orbital energy of the galaxy. The timescale for the galaxy to sink to the center is given by

$$t_{\text{sink}} = \frac{r^2 V_c}{Gm} \frac{1}{\mathcal{F} \ln \Lambda} \tag{6.4}$$

where V_c is the circular velocity of the halo and r the initial radius of the satellite of mass m. $\mathcal{F}$ and $\ln \Lambda$ are dimensionless constants relating to the velocity anisotropy of the dark matter particles in the halo and the relative range of encounter scales (Chandrasekhar 1943). The satellite's dark matter halo also looses mass as it spirals in, tending to increase the dynamical friction timescale. These uncertainties can be calibrated by numerical simulations (Boylan-Kolchin et al. 2008).

Dynamical friction leads to galaxies spiraling into the center of the cluster. Indeed, the central galaxies of clusters have unusual properties, not seeming to fitting extrapolation of the normal luminosity function and reflecting a flattening of the galaxy mass-metalicity relation. When clusters merge, dynamical friction will cause the central galaxy of each system to merge together. This is a likely explanation of clusters, like the Coma cluster, that contain two (or more) dominant galaxies.

2.5.3 Ram-Pressure Stripping

Cluster galaxies are moving through the diffuse intra-cluster medium (ICM) and are subject to hydrodynamic forces. In particular, if the galaxy contains a disk of cold gas, it must force the ICM aside for the galaxy to pass through. If the gravitational force holding the gas in the plane of the galaxy is too weak, the gas will be stripped. This condition can be written as

$$\rho_{\text{ICM}} V^2 > 2\pi G \Sigma_*(r) \Sigma_g(r) \tag{6.5}$$

where ρ_{ICM} is the density of the diffuse external gas V in the orbital velocity of the galaxy, $\Sigma_*(r)$ is the stellar surface density of the galaxies disk at radius r, and Σ_g the surface density of gas at this radius. If the condition is satisfied, gas will be stripped from this radius. If the

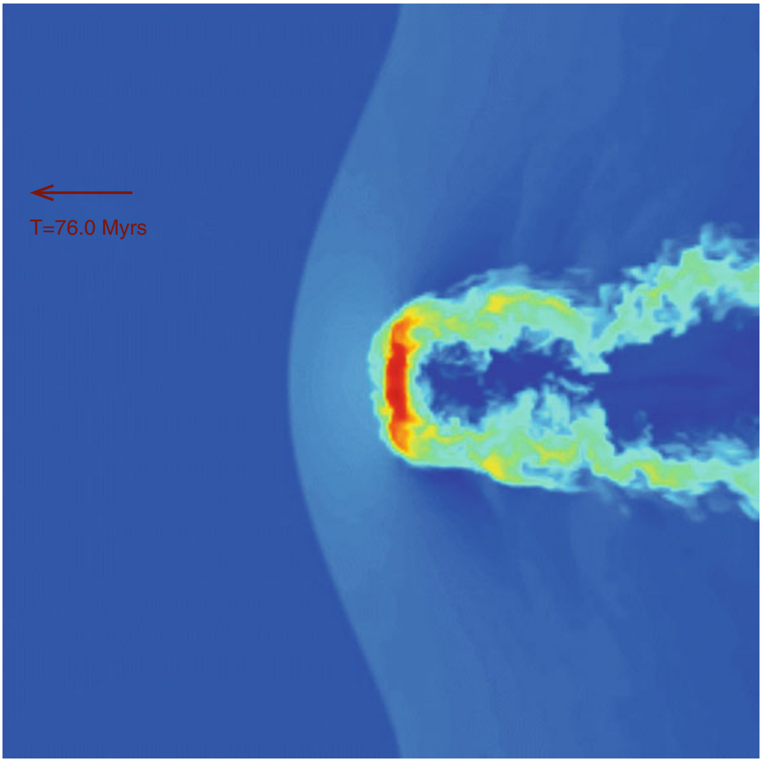


Fig. 6-5
A simulation of the effect of ram-pressure stripping on a spiral galaxy. As the galaxy passes through the intra-cluster medium, the ram pressure sweeps material out of its disk leaving a trail of ionized material behind it. Such trails have been observed behind spiral galaxies in clusters (Image credit: Quilis et al. 2001)

galaxy is moving sufficiently fast, all the gas in the galaxy will be removed. At lower velocity, only material in the outer parts is removed (Gunn and Gott 1972). The results of these simple analytic arguments can be verified with numerical simulations (➋ *Fig. 6-5*). These show that the formulae are broadly correct (Quilis et al. 2001), but highlight a number of potential issues such as the compression of the interstellar medium during the stripping process.

The ram-pressure stripping mechanism presents a plausible explanation of the absence of conventional spiral galaxies in clusters. Indeed, some examples of galaxies in the process of being stripped have been clearly identified through high-resolution ultraviolet imaging of galaxies in nearby clusters. However, the stripping process is only likely to be strong if the motion of the galaxy is rapid and the ICM is sufficiently dense. Since lower mass systems, such as galaxy groups, contain a much lower density plasma, this is not likely to explain the relative abundance of the early-type galaxies in groups. Yet, observations of galaxy groups show that the passive galaxy sequence and many of the characteristic properties of cluster galaxies are already established in lower mass groups. It seems that this mechanism cannot be the dominant driver of galaxy ecology.

2.5.4 Strangulation

This is a theorist's mechanism, and not directly observed. Galaxy formation models suggest that galaxies continually recycle gas between their disk and halo (White and Frenk 1991). Feedback from supernovae ejects the gas from the disk regulating the star formation rate of the galaxy. Without this process, galaxy formation is far too efficient to match the very-low-observed stellar mass fraction of the universe. (Clusters are a good census: only 10% of the cluster baryons are locked into stars.)

In cluster galaxies, this cycle is easily interrupted by ram-pressure stripping of the loosely attached material in the halo of the galaxy. This is a distinct process from the stripping of the tightly bound material in the galaxy disk. However, the process can be described by the same formula, but need to reassess the coefficient because of the relative distributions of mass and gas. Essentially, the same formula applies as for the ram-pressure stripping of disk gas but with a new coefficient to allow for the revised geometry (McCarthy et al. 2008).

This process is effective even in small groups. Indeed, it seems to be too effective when incorporated into a cosmological self-consistent model for galaxy formation and evolution. The current challenge is to realistically incorporate the mechanism into cosmological galaxy formation models. Current schemes seem to predict that the mechanism is rather too effective compared to the observational data.

2.6 Evolution of Galaxy Clusters

The idea is appealing and simple. In order to map out the formation of cluster galaxies, we should find clusters at high redshift and compare them to local systems. The process has been likened to understanding the family history through a set of snapshot photographs. Of course, it is not quite so simple since we cannot observed the same cluster at different moments in its history, and we must piece the story together from statistical arguments.

It therefore needs care: (1) we need to match the masses of galaxies, not their luminosities, and to be careful to make measurements the rest frame of each cluster. (2) Massive clusters at high redshift will grow in mass by the present-day. They are also so rare that it can be problematic to sample a sufficiently large volume in the local universe. (3) We must pay attention to how the clusters are selected. If selection is based on an observed-frame optical band, high-z clusters will be preferentially selected if they contain more blue galaxies.

2.6.1 Color Evolution

One of the most striking results from the early observations of clusters was the Butcher–Oemler effect (Butcher and Oemler 1984). The authors reported the colors of galaxies in optically selected distant clusters. To the surprise of the community, they discovered a rapidly rising fraction of blue galaxies in the clusters. This was unexpected in that the effect was seen out to quite moderate redshifts ($z = 0.5$), in stark contrast to the paradigm of the time that cluster galaxies were old, with ages comparable to that of the universe.

The result is illustrated in ➋ *Fig. 6-6* for a modern sample. The line illustrates the evolution in the fraction of star-forming galaxies based on the sample of clusters at $z < 0.3$. If the trend could be extrapolated, then there would be almost no red (passive) galaxies in clusters at $z \sim 1$.

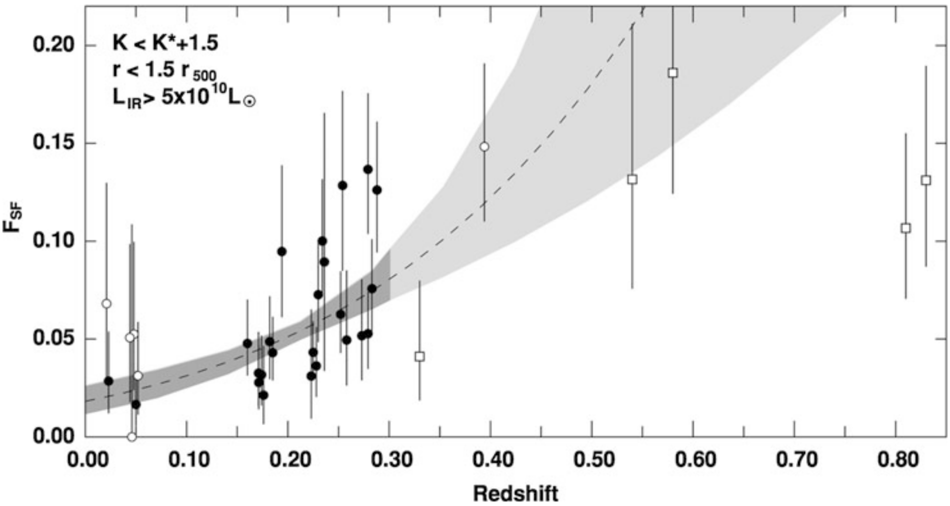


Fig. 6-6

A modern measurement of the Butcher–Oemler effect. The Butcher–Oemler effect was originally seen as an increase in the fraction of blue (and hence star forming) galaxies in clusters towards higher redshifts. Here, the same effect is seen in a sample of carefully selected clusters using mid-infrared observations to identify star-forming galaxies. A rapid increase with redshift is suggested (*dashed line*, fitted to data $z < 0.3$), but scatter between clusters is very large. *Solid symbols* show results from high X-ray luminosity clusters, while *open symbols* denote systems of lower X-ray luminosity. *Open squares* show higher redshift clusters observed by Saintonge et al. (2008). Clearly, these clusters do not follow the trend suggested by the *dashed line* (Image credit: Haines et al. 2009)

Butcher–Oemler's original result has been heavily criticized. Clearly, not all the clusters follow the same trends, and some local but less relaxed clusters have blue fractions comparable to clusters at higher redshifts. Moreover, Butcher and Oemler's original selection of the clusters was inhomogeneous, with some of the clusters being selected from blue photographic plates. Clearly, this could bias the cluster sample to systems that contained more blue galaxies than the ensemble average. Another important bias is that the greater star formation rates of high-redshift field galaxies, combined with the bluer rest-frame pass bands typically used, mean that magnitude limited samples typically contain disproportionately many dwarf galaxies compared to local systems.

However, these biases do not seem to be sufficient to fully account for the effects, and the result has been reported in more modern cluster samples including results based on mid-IR star formation rates, as shown in the figure. It seems that these biases cannot fully explain the effect.

However, another bias is much harder to remove. Because the universe is younger (and the masses of clusters are more extreme compared to the average halo mass), the growth rates of clusters are significantly higher in the past compared to the present-day. Infalling galaxies cannot easily be distinguished from the virialized cluster population. Moreover, the field galaxy population is much more active at higher redshifts, with a larger fraction of the population having blue colors. The situation is therefore unclear, and a modern interpretation could be that the increase in active galaxies results from the changes in the field galaxy population rather than from changes in the physics that occur within galaxy clusters.

2.6.2 E+A Galaxies

In the classic literature, the discussion of the evolution of cluster galaxies is closely related to the presence of a "new" population of "post-starburst" galaxies in distant clusters. These galaxies are often referred to as "E+A" galaxies because their distinguishing feature is spectra that are dominated by strong Balmer absorption, most notably the Hδ line. The line strengths, combined with the absence of emission lines, seen in some spectra cannot be reproduced with declining star formation history. The best described as the spectrum of an A star superposed on that of an elliptical galaxy (hence the tag "E+A") and required a strong burst of star formation followed by an abrupt truncation (Couch and Sharples 1987). Recently, it has been realized that an alternative interpretation is possible in which the star formation is ongoing but heavily dust obscured. Only the A star population escapes from the strong dust obscuration creating the strong absorption.

The E+A population is now realized not to be confined to clusters, and examples of the E+A type are evident in the local galaxy populations where they can be observed in detail to explore the cause of the unusual activity. It is likely that the population originally identified in distant clusters is the result of the greater star formation rates of field galaxies and the subsequent suppression of their star formation in the cluster or galaxy group environment.

2.6.3 Cluster Archaeology

An alternative approach to cluster evolution is to observe the properties of local galaxies in detail. High signal-to-noise observations of galaxies in local clusters allow the degeneracy between metalicity and average stellar age to be broken. This work requires great care and the development of spectral models that allow for nonsolar element abundances.

While these results show that the high-mass cluster galaxies are old, as expected, they reveal a significant trend for the lower mass galaxies to be younger (➋ *Fig. 6-7*). This is an intriguing result that seems to require a degree of co-ordination between the metalicity of the lower mass galaxies and their weighted age. This cancelation does not arise naturally in theoretical models that have efficient feedback.

2.6.4 The Luminosity Function

A fourth approach to the evolution of cluster galaxies is to observed the system luminosity function. This requires care since the normalization of the luminosity (or mass function) depends on the overall mass of the halo and not the volume surveyed. To compare, the mass of the halo requires an independent measure of the system mass, such as that obtained from gravitational lensing.

However, the relative shapes of the mass function can be compared much more simply. In particular, the fraction of the total stellar mass contributed by high, and low-mass passive

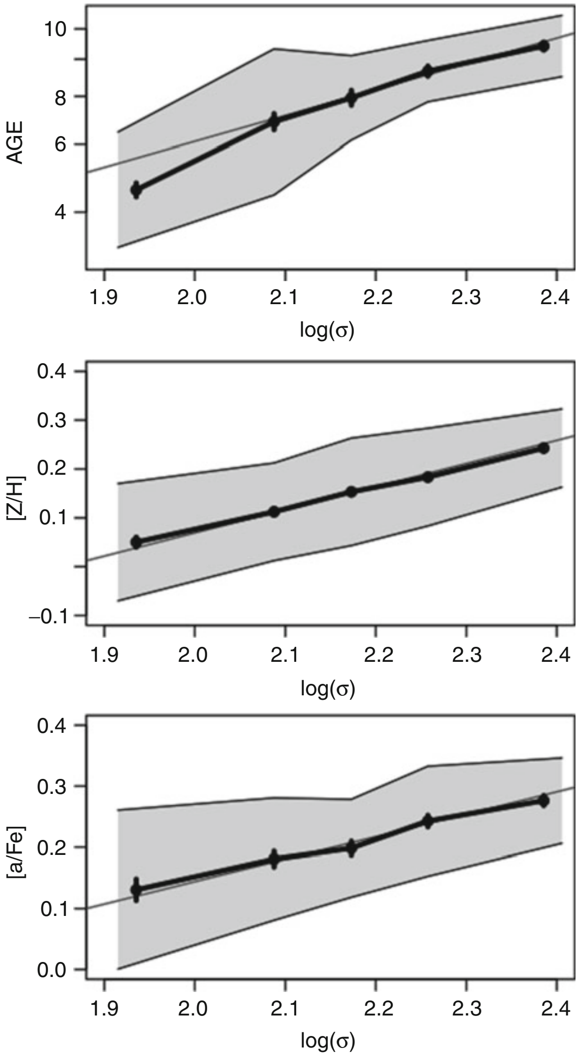


Fig. 6-7

The dependence of age and metal abundance on galaxy velocity dispersion (effectively system mass) for a large sample of nearby early-type galaxies. The *thick line* shows the average age, metalicity, and α element enhancement. The *grey-shaded* region indicates the intrinsic scatter about the mean trend (Image credit: Nelan et al. 2005)

galaxies has proved to be a useful diagnostic and one that can be robustly measured. De Lucia et al. (2004) reported a relative porosity of low-mass red-sequence galaxies on the basis of HST imaging. The result has been confirmed in several (although not all) subsequent studies, such as that shown in *Fig. 6-8*.

It is interesting to compare this with the emergence of relatively young ages for low-mass galaxies from cluster archaeology. A simple interpretation suggests that the correlation between

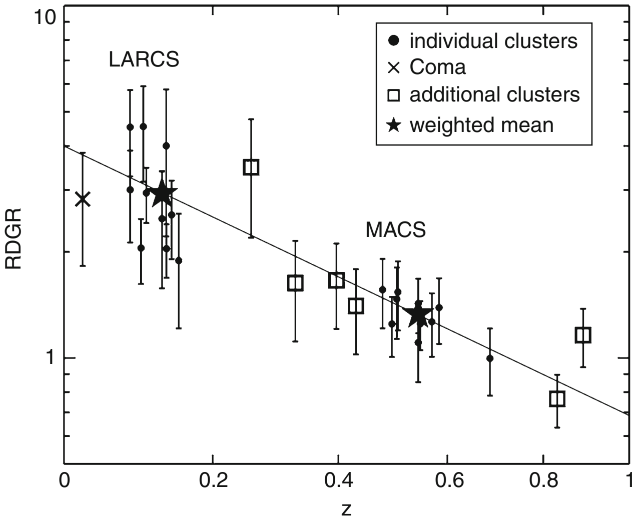


Fig. 6-8
The evolution of the ratio of dwarf (fainter than $M_V = 20$) to giant (brighter than $M_V = 20$) red-sequence galaxies as a function of redshift. The sample shown here is taken from Stott et al. (2007) and uses carefully matched samples of X-ray-selected clusters. A strong trend is seen with fewer red-sequence dwarf observed in the higher redshift clusters. This suggests that star formation in faint galaxies has not yet been suppressed in these systems

age and stellar mass should twist the color-magnitude sequence, increasing its slope at higher redshift (in contrast to the mild slope evolution that is observed). However, this is based on the assumption that the star formation rate slowly declines as the galaxy is incorporated into the clusters. A more complex interpretation appears to be required, in which galaxies make a rapid switch between the passive and star-forming sequences. It is interesting to note that the "rapid switch" picture is supported by the strong presence of two sequences even in studies of the group and field population (Balogh et al. 2004).

The discussion of the evolution of the stellar mass function has highlighted the need for care when comparing clusters at different redshifts and even between systems of different mass. It is now very apparent that stellar masses need to be compared on a like-for-like basis and that samples need to be compared on the basis of stellar mass, not luminosity (particularly luminosity in bluer bands). This not only complicates the comparison of systems at different redshift, but it makes it harder to compare systems of different total halo mass (since this also affects the stellar mass function).

2.6.5 Morphology

The launch of the HST has offered the possibility of studying the morphologies of galaxies in distant clusters, allowing this to be compared to the changes in the star-forming properties. This topic has resulted in extensive controversy largely because of the difficulty in obtaining objective computer-derived morphologies. Carefully moderated "eyeball" morphologies consistently suggest that distant clusters contain fewer S0 galaxies and more spiral galaxies than local counterpart clusters (Dressler et al. 1997). Interestingly, the clusters contain similar fractions of early-type galaxies at all redshifts, hinting that the cluster population consists of an intrinsic elliptical galaxy population and an S0 population that has built up over time. It is difficult to reproduce this result, with current galaxy formation models.

The spiral to S0 transformation would appear to reinforce the evolution of galaxy colors first suggested by the Butcher–Oemler effect. However, the same caveats about the role of magnitude limits, cluster selection, and galaxy infall apply equally here. Nevertheless, many authors are convinced of the evidence for a long timescale of transformation of spiral galaxies to S0, others identify galaxy groups as the key environment in the history. Care, however, needs to be exercised since many studies average the whole galaxy population together and do not fully account for the biases in the stellar mass function of cluster, group, and isolated galaxies.

2.6.6 Other Wavebands

Observations in the FIR have opened up the possibility of including deeply dust-obscured star formation in the census of star formation in clusters and groups of galaxies. In particular, the observations stand to reveal a source of star formation in the E+A galaxy population.

To date, the results have been mixed. Some authors report the discovery of strong FIR sources in clusters, consistent with such deeply obscured star formation (Geach et al. 2009). However, the sources appear to be rare, and when the mass biases are taken into account, it is unclear whether the population is specific to the cluster environment, or is better accounted for by the enhanced infall of cluster galaxies. With the advent of Herschel and better models for the evolution of galaxies, this is a rapidly advancing field.

2.7 The Relation of Clusters to Galaxy Groups

The above discussion makes it clear that the evolution of galaxy clusters cannot be separated from the evolution of galaxy groups and a great deal of effort is being targeted on measuring the properties of galaxies in less rich environments, or even to measure the properties of galaxies as a function of their clustering strength rather than assigning individual halo masses to galaxies. The later approach has the advantage that it can be computed using objectively defined statistical measures such as the correlation function or the marked correlation function.

Studies of galaxy groups are, however, hard work compared to studies of clusters since targetted spectroscopic campaigns return a much larger fraction of field galaxies compared to group members. One of the best approaches is therefore to undertake large, highly complete redshift surveys (such as the zCOSMOS survey (Maier et al. 2009)) and to select galaxy groups from the survey. These can then be followed up in more detail. The results are exciting – at intermediate redshift, the galaxy groups are offset from the field but have a larger proportion of active galaxies than similarly selected groups at low redshift. These results are now being confirmed out to higher redshifts. At $z = 1$ groups from the zCOSMOS survey appear to be dominated by a transition population of E+A galaxies (Balogh et al. 2011). These results have encouraged many authors to consider that galaxy groups are the powerhouse of galaxy transformation.

Another approach to this issue is to assign each galaxy an environment based on its local density. Peng et al. (2010) use this approach to present an interesting view of the relation between galaxies and environment. Presenting the results in this was, they show that the fraction of star-forming galaxies can be effectively written as the product of two quenching terms, one based on the galaxy's mass and the other on its local density. The two terms are independent, which seems to agree well with the expectation of theoretical models. This will be an interesting angle to pursue with new, even larger, redshift surveys.

2.8 Intra-cluster Light

In a later section, I will present the role of dynamical friction in causing the orbits of galaxies to decay. As the orbit decays, the galaxy becomes subject to stronger tidal forces which may strip stars from the outer parts of the galaxy. The stars that are stripped in this way would continue to orbit in the cluster creating a diffuse interstellar glow. Although faint, the diffuse light is just observable (Gonzalez et al. 2005), particularly around the dominant galaxy where it is concentrated to the center of the gravitational potential.

Although the existence of the intra-cluster light is now well established, The total fraction of stars in this diffuse form is still a matter of controversy. It is difficult to make this assessment directly because of the surface brightness that drops rapidly away from the central galaxy. McGee and Balogh (2010) circumvent this difficulty by measuring the intergalactic supernova rate in clusters. Current estimates suggest that the intra-cluster light in clusters accounts for about 10% up to 50% of the total star light. While recent measurements in lower-mass haloes have suggested a larger contribution (based on extrapolation to low surface brightness), it is hard to see how these results can be reconciled with the hierarchical growth of massive clusters from lower mass systems.

3 X-Ray Emission

3.1 The Physics of X-Ray Emission

The optical properties of galaxy clusters reflect only a small fraction of their baryonic mass. Most of the baryons in a cluster are in the form of a diffuse intra-cluster medium (ICM). This relatively dense, high-temperature plasma gives rise to copious X-ray emission. The typical temperature of the cluster plasma is 10^7–10^8 K, and the core density reaches 0.01–0.1 atoms per cm^{-3}. At these temperatures, the plasma is highly ionized leading to X-ray emission from thermal Bremsstrahlung (also referred to as free–free emission) and inner-shell electron capture and transitions.

3.1.1 Thermal Bremsstrahlung Emission

Conceptually, thermal Bremsstrahlung is straightforward. It arises as the trajectories of electrons are deflected by the strong fields of ionized nuclei, including H^+ and He^{++}. The acceleration of the electron generates the emission of a photon. Computation of the actual spectrum involves integration over impact parameters and the thermal distribution of electron velocities. For an ion with charge Z and a plasma with electron temperature T_e and electron and ion densities n_e and n_i, respectively, the emissivity (per unit volume per unit frequency) is given by

$$\epsilon_Z^{ff}(\nu) = A_Z n_e n_i T_e^{-1/2} \exp(-h\nu/k_B T_e) \tag{6.6}$$

where k_B is the Boltzmann constant, h Plank's constant, and the normalization constant is given by

$$A_Z = \frac{2^5 \pi e^6}{3 m_e c^3} \left(\frac{2\pi}{3 m_e k_B} \right)^{1/2} Z^2 g_{ff}(Z, T_e, \nu) \tag{6.7}$$

where m_e is the electron mass, e the electron charge, and g^{ff} the Gaunt factor, a slowly varying function that corrects quantum mechanical effects (Sarazin 1988).

The spectrum has a continuous exponential distribution. If plotted in terms of $\log(\nu)$, this translates to a break at high energies corresponding to the thermal temperature of the plasma. In order to obtain the total luminosity arising from free–free emission, we must integrate over frequency and the volume, V, of the cluster and sum over each species:

$$L^{ff} = \sum_Z \int_\nu \int_V \epsilon_Z^{ff}(\nu)\, d\nu\, dV \tag{6.8}$$

Although the emission formula appears complex, it is characterized by its dependence on the square of the density of plasma and its $T_e^{1/2}$ temperature dependence. The later arises because a higher temperature pushes the photon energy cutoff to higher energy ($\propto T_e$), while the higher temperature reduces the effectiveness of individual collisions ($\propto T_e^{-1/2}$). Free–free emission in spectrum is dominated by collisions with hydrogen and helium nuclei.

The dependence on the square of the plasma density is a key result. This has important consequences for the cooling instability since most heating mechanisms depend linearly on density. This makes it difficult for a heating mechanism to simultaneously compensate for radiative cooling over a range of radii.

Thermal Bremsstrahlung dominates the emissivity at temperatures above 8×10^7 K, where the common elements become completely ionized. At lower temperatures, electron energies are comparable to the ionization potential and free-bound and bound–bound transitions become important.

3.1.2 Bound–Bound Electron Transitions

X-ray emission also arises from bound transitions and the capture of electrons into bound shells. Typically, the electron will be captured into a high-energy state, and a cascade of X-ray transitions will result. Alternatively, one of the ion's inner-shell electrons will be excited by collisional processes. The resulting spectrum of emission lines depends both on the temperature of the plasma and the abundance of different species. The calculation is complex because it requires careful determination of the ionization balance of each species. More details of the calculation are given in Sutherland and Dopita (1995). Recent calculations of the emission spectrum are discussed in Wiersma et al. (2009), where the role of photo-ionization is considered in depth. Fortunately, this correction is less important for high-temperature plasmas such as those in clusters. Line emission dominates the emissivity at temperatures below 10^7 K.

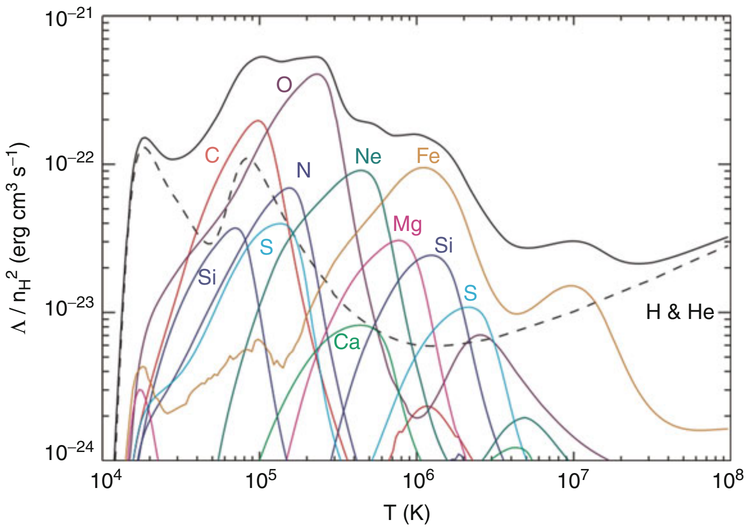


Fig. 6-9

The contribution of different elements to the overall cooling rate of the intra-cluster plasma. The plot here illustrates the temperature dependence of the total emissivity for a solar abundance plasma (*solid lines*). The contributions of different elements are indicated. The emissivity of a primordial plasma containing on hydrogen and helium is indicated by a solid line. These calculations assume that the plasma is in collisional ionization equilibrium, and at low temperatures, the cooling curve is modified substantially in the presence of an ionizing background (Image credit: Wiersma et al. 2009)

Just as for thermal Bremsstrahlung, the total emissivity depends on the square of the plasma density since ionization mechanism is (at cluster temperatures) a two-body process. Moreover, the shape of the emission spectrum can be used to infer the plasma temperature and the abundance makeup. If the relative abundance of the elements is known, this makes it possible to accurately infer the density and temperature of the plasma even in relatively low-quality X-ray data (➲ *Fig. 6-9*).

It is interesting to consider the temperature dependence of the total (or volumetric) emissivity in more detail. For a plasma that has not been enriched by stellar nucleosynthesis, the temperature dependence of the total emissivity is dominated by two peaks corresponding to the ionization potentials of hydrogen and helium. As the metal abundance of the plasma increases, the elements C, O, and Ne play an important role. As a result, the total emissivity of the plasma has a very strong dependence on metal enrichment in the temperature range 10^5–10^7 K. This has important consequences for the predicted luminosities of groups and clusters. Cross talk between the formation of stars and the cooling of gas in galaxy clusters makes simulations of the universe extremely challenging.

In addition, nonthermal processes can also lead to the emission of higher energy photons. The comptonization of the cosmic microwave background, in particularly, is potentially important at higher redshift.

3.1.3 Total Emissivity

The total (or volumetric) emissivity is conveniently written as

$$\epsilon = n_H^2 \Lambda(T, Z) \tag{6.9}$$

where n_H is the hydrogen number density and Λ encapsulates the complex function of temperature and ion abundance. This must be computed with computer codes such as those of Sutherland and Dopita (1995) and Wiersma et al. (2009). Tabulated results are readily available, but care needs to be taken to ensure that the density term in (➔ 6.9) is consistently defined; many authors use a pre-factor of $n_e n_H$. For a primordial plasma containing 25% helium by mass, $n_e = 1.167 n_H$. Fortunately, the gross behavior of the emissivity can be approximated by

$$\Lambda \approx 3 \times 10^{-23} (T_8^{1/2} + 0.5 f_m T_8^{-1/2})\ \mathrm{erg\,cm^3\,s^{-1}} \tag{6.10}$$

where T_8 is the plasma temperature in units of 10^8 K and f_m is a factor that takes into account the metal abundance of the plasma (Peacock 1999). For solar abundances, $f_m = 1$, and for a primordial plasma, $f_m \approx 0.03$. Note that at high temperatures, we recover the $T^{1/2}$ dependence expected for thermal Bremsstrahlung, while at temperatures lower than those of groups and clusters, the dependence is $T^{-1/2}$.

3.2 The Baryon Content of Galaxy Clusters

We can apply our understanding of the emission mechanisms to determine the hot gas content of galaxy clusters. The temperature can be inferred from the spectrum. If consideration is restricted to fixed element abundance ratios, this can be achieved even at relatively low spectral resolution. This makes it possible to construct detailed temperature (and abundance) maps for nearby clusters.

3.2.1 Temperature and Density Profiles

The temperature profile is close to isothermal, with most clusters showing a modest decrease in temperature towards their center, and a slight drop in temperature towards their outer edge. The temperature of the system can be estimated from the Virialized mass, M_v, of the object

$$k_B T = \frac{1}{2} \mu m_p V_c^2 \tag{6.11}$$

where μ is the average particle mass (0.59 for a primordial plasma), m_p is the proton mass, and V_c is the circular velocity of the halo, defined as

$$V_c = G^{1/2} M_v^{1/3} \left(\frac{4\pi}{3} \rho_v \right)^{1/6} \tag{6.12}$$

where ρ_v is the average density of the halo. Typically, we assume $\rho_v = 200 \rho_{\mathrm{crit}}$, where ρ_{crit} is the critical density of the universe at the redshift of collapse. Combining these expressions, we expect a mass–temperature relation close to

$$T \approx 1.6 \times 10^7 \left(\frac{M_v}{10^{14}\, M_\odot} \right)^{2/3} \mathrm{K} \tag{6.13}$$

This is within 15% of the observed relation (Vikhlinin et al. 2006).

The density can be reconstructed from the observed X-ray surface brightness. Of course, care is needed to reconstruct the 3-dimensional density distribution from the 2-dimensional measurements. This typically involves making an assumption of spherical symmetry which allows the contribution from successive shells to be determined.

Clusters are conventionally fitted by a density profile of the form

$$\rho = \frac{\rho_c}{(1 + (r/r_c)^2)^{\frac{3}{2}\beta}} \tag{6.14}$$

where ρ_c, r_c, and β are fitting constants. Typically, $\beta = 2/3$ provides an adequate fit although more recent papers fit more complex profiles, or dispense with simple fitting formulae altogether and use the results of numerical simulations directly (Vikhlinin et al. 2006). Central gas densities greater than 10^{-2} atoms per cm^3 ($\rho_g \sim 3 \times 10^{14}\, M_\odot\, \mathrm{Mpc}^{-3}$) and core radii around 100 kpc are typical. The plasma is typically enriched with metals, typically to about 1/2 of the solar abundance (Leccardi and Molendi 2008).

Integrating the profile gives the total gas mass. This is typically 12% of the total system mass (see below for a discussion of techniques for estimating the total mass of the system). The gas content of clusters is much greater than the mass in stars and in cold gas, which makes up only 10% of the total baryon content. A great deal of the residual uncertainty comes from the diffuse intra-cluster stars.

These fractions are very comparable to the stellar fraction of the universe as a whole, with the notable exception that it is not possible to directly detect intergalactic plasma because its temperature is too low for efficient X-ray emission and its presence is inferred indirectly from cosmological measurements of Ω_b/Ω_m.

3.2.2 The Entropy Distribution

Measurement of the temperatures and density of the intra-cluster plasma may be summarized in terms of the "entropy" of the plasma at each radius. Typically, X-ray astronomers use the adiabat of the gas

$$K \equiv k_B T \rho^{-2/3}$$

as a proxy for the entropy (thermodynamic entropy is proportional to $\log K$). The adiabat is a useful quantity, because it does not change as the gas is adiabatically compressed due to changes in the gravitational potential. Moreover, the buoyancy of the gas results in a natural sorting of the material such that the lower entropy material sinks to the center of the cluster, while the higher entropy material rises to the outside. Numerical experiments show that this segregation occurs rapidly as the cluster grows. Hydrostatic equilibrium, buoyancy, and an insensitivity to the outer boundary conditions make it possible to uniquely specify the gas distribution in the cluster in terms of the gravitational potential and the entropy distribution.

Entropy also provides a good variable for cluster studies since it can be used to infer the total heating and cooling of the gas as the system has formed. Heating sources include the shock heating of gas as it enters the cluster (and energy injected by a central radio galaxy), while cooling is dominated by the radiative losses of the ICM. These ideas can be used to treat clusters

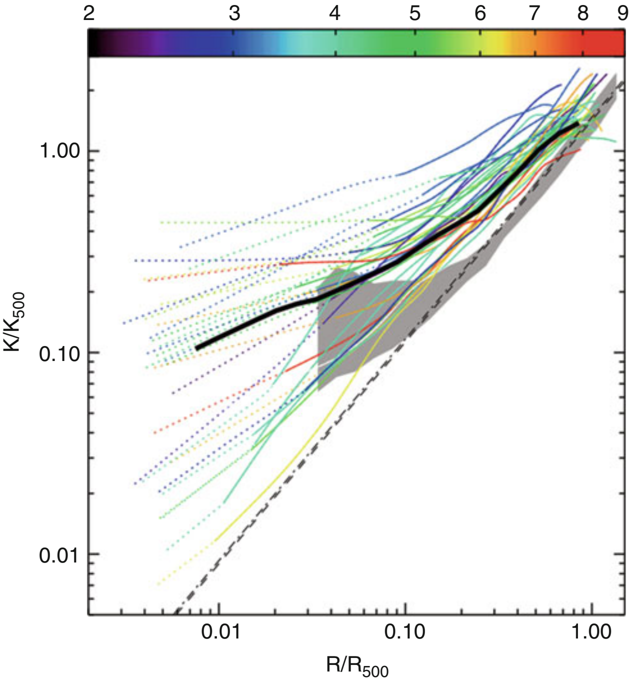


Fig. 6-10

The entropy profiles of intra-cluster medium in galaxy clusters taken from the REXCESS survey. The entropy profile in each cluster has been scaled by dividing by the characteristic entropy expected for a cluster of the observed temperature $\left(K_{500} = \frac{1}{2}\left(\frac{2\pi G^2 M_{200}}{15 f_b H(z)}\right)\right)$. The colors of the *lines* indicate the cluster temperature. *Black lines* and *shaded regions* show the expectations of theoretical models. Scaling by K_{500} brings the profile to similar values at the outer edge of the system, but the profiles have a wide range of slopes. Typically, the lower temperature clusters have shallower profiles. This results from a loss of the lower entropy material from these systems (Image credit: Pratt et al. 2010)

of galaxies as "cosmic calorimeters" allowing us to gain great insight into the thermal history of the material that is left over from the process of galaxy formation (*Fig. 6-10*).

Observations of the entropy distribution functions of clusters of different mass reveal an unexpected scaling. Rather than the entropy distributions being scaled copies of each other, gas in lower mass systems has a much flatter distribution of entropy. This leads to a much lower concentration of the central gas and hence to much weaker X-ray emissivities of low-mass clusters than expected. The difference cannot be attributed to a greater stellar fraction in the lower mass systems, but the distinction is clearly connected to the "cooling flow paradox" that we discuss below. Initially, the differences in entropy profile were thought to result from a minimum entropy boost associated with an early epoch of galaxy formation (Ponman et al. 1999), but it now seems more appropriate to view this as arising from the loss of low entropy material from lower mass groups during the cluster formation (McCarthy et al. 2011). This has provided powerful but complimentary evidence for the role of AGN in regulating the formation of galaxies.

3.3 The Cooling Instability

A useful quantity is to define the cooling time of the X-ray plasma as the ratio of its thermal energy to its radiation rate:

$$t_{\mathrm{cool}} \equiv \frac{\frac{3}{2} n\, k_B T}{n_H^2 \Lambda} \tag{6.15}$$

where n is the total number of thermal particles in the plasma and n_e and n_H are the electron and hydrogen number density of the plasma. For a primordial plasma containing 25% helium by mass, $n = \frac{\rho_g}{0.59 m_p}$ and $n_H = \frac{\rho_g}{1.33 m_p}$, where m_p is the proton mass and ρ_g is the plasma density. Using the approximation for Λ introduced previously (❯ 6.10), one obtains

$$t_{\mathrm{cool}} \sim 16 \left(\frac{\rho_g}{(10^{14}\, M_\odot\, \mathrm{Mpc}^{-3})} \right)^{-1} \left(T_8^{-1/2} + 0.5 f_m T_8^{-3/2} \right)^{-1} \mathrm{Gyr} \tag{6.16}$$

Inserting a typical core gas density and temperature gives $t_{\mathrm{cool}} \sim 3\,\mathrm{Gyr}$, significantly shorter than the age of the universe.

The key point is that denser plasma has a shorter cooling time. Reducing the temperature also shortens the cooling time. This leads to a cooling instability. As the plasma radiates, it is squeezed by the weight of the material at larger radii. The net effect is that the pressure remains roughly constant, but the density rises. The rise in density further increases the emission rate, which in turn accelerates the increase in density. This leads to a radial inflow of gas referred to as a "cooling flow" (Fabian 1994). Initially, the rate of flow is much smaller than the system sound speed, and it is adequate to consider the cluster proceeding through a sequence of quasi-hydrostatic states. Eventually, this approximation breaks down, and the plasma becomes subject to local instabilities that grow on a timescale comparable to the free-fall time of the gas.

Clusters of galaxies are thus intrinsically unstable (❯ *Fig. 6-11*), leading to the expectation that the observed cluster should contain a net inflow of material towards the center of the gravitational potential. Fortunately, the timescale for such flows in clusters is relatively long compared to the overall age of the universe. Nevertheless, while we would not expect the cluster to cool completely, a significant fraction of the material is expected to flow towards the center of the system. Paradoxically, there is no evidence for such "cooling flows" on the scale expected.

A subtle but important point is that the cooling of the plasma does not necessarily result in a drop in temperature if the cooling time is longer than the system dynamical time. In a quasi-hydrostatic, flow the temperature reflects the shape of the potential. However, once the gas is sufficiently dense, the cooling time becomes short compared to the dynamical time and clumps of gas will condense out of the flow and cool.

In contrast to the situation in galaxy clusters, the cooling time is relatively short in galaxy systems. A relevant comparison is with the local dynamical time:

$$t_{\mathrm{dyn}} = (G\rho_{\mathrm{tot}})^{-1/2} \tag{6.17}$$

where ρ_{tot} is the total density (including both dark matter and gas). Assuming $\Omega_b/\Omega_m = 0.167$ and that dark matter and gas have the same distribution (a poor approximation within clusters!), we can write

$$t_{\mathrm{dyn}} \approx 2 \left(\frac{\rho_g}{10^{14}\, M_\odot\, \mathrm{Mpc}^{-3}} \right)^{-1/2} \mathrm{Gyr} \tag{6.18}$$

Comparing this with (❯ 6.16), $t_{\mathrm{cool}}/t_{\mathrm{dyn}} >> 1$ in the centers of galaxy clusters. However, the ratio rapidly declines as the virial temperature of the halo falls, and the way in which cooling proceeds

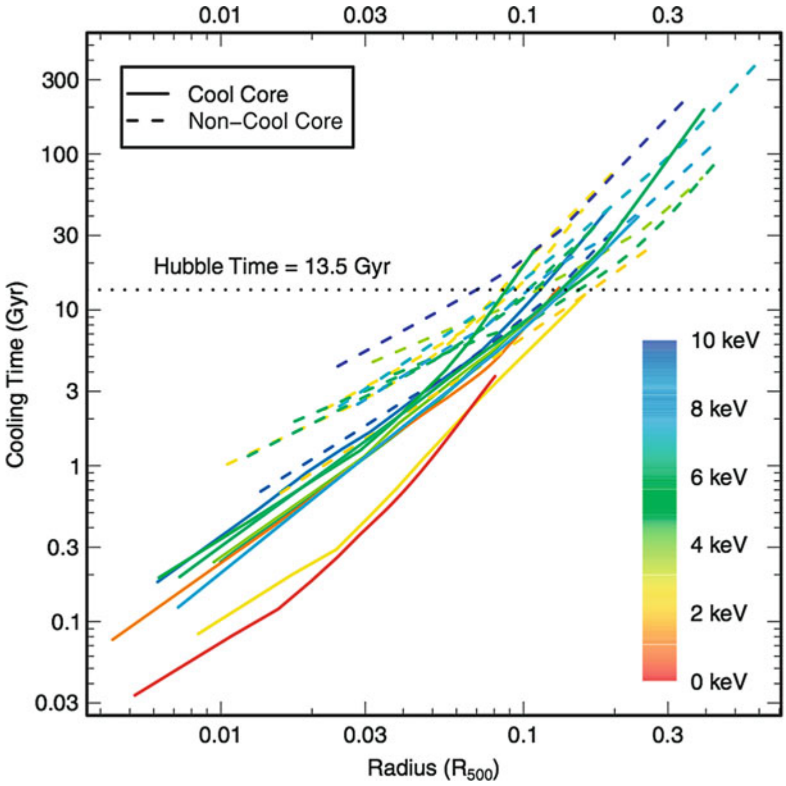


Fig. 6-11

The cooling time profiles of a sample of X-ray clusters taken from Sanderson et al. (2006). The cooling time of the plasma decreases strongly with decreasing radius so that the central cooling time of most clusters is much less than the age of the universe. *Solid* and *dashed lines* distinguish between clusters that have a central dip in their temperature profile and those that do not (Image credit: Sanderson et al. 2006)

in galaxy-scale haloes is potentially quite different to that in higher mass cluster systems (White and Frenk 1991; Dekel and Birnboim 2006). This leads to an important distinction between hot-mode accretion (i.e., cooling flows) and the "cold-mode" accretion that dominates the growth of galaxies. In galaxies, the fueling rate is therefore determined by the rate of growth of the halo and not the rate at which gas is able to radiate its energy. The accretion rate of galaxies is thus a strong function of redshift.

Is strong AGN feedback in clusters inevitable, or should we be surprised by the frequency and strength of the energy input? This is currently a subject of much debate. On the one hand, if the AGN were not present or effective, this would lead to a pileup of cold gas at the center of the cluster. Surely, this would eventually lose its angular momentum and find a way to accrete onto the central black hole and thus establish a regulating feedback loop! But while this picture is appealing, there are many missing pieces. For example, why are the processes so effective at transporting the cooling gas down to a few Au of the black hole? What would happen if the accretion rate were so high that the central accretion disk radiated the accretion energy away rather than producing a powerful radio jet? Hopefully, we will be able to piece together a much more complete picture in the near future.

3.3.1 Cooling Flows: Comparison to Observations

Detailed calculations of the cooling time profiles of galaxy clusters suggest that material should be flowing to the center at typical flow rates of 100–1,000 $M_{\odot}$year^{-1} (Peres et al. 1998). The rates are sensitive to the resolution of the X-ray observations (higher flow rates are found in better resolved systems), and samples are biased by X-ray selection (since clusters with higher cooling rates also have higher X-ray luminosity). Nevertheless, this work leads to the expectation that, in the absence of an effective heating mechanism, mass should be transported to the center of the cluster, cool out and form into stars. Some modern determinations of cooling time profiles are shown in ➲ *Fig. 6-11*.

However, the short cooling times seem contrary to observations.

(1) Although central cluster galaxies often contain significant populations of young stars and cold gas, the inferred star formation rates are roughly 1/10 of that inferred from the analysis above. Similarly, the observation of extended Hα emission filaments does not directly imply the high star formation rates since the emission strength greatly exceeds that expected even under the flow scenarios above. It seems more likely that the filamentary emission is indicative of the cosmic ray flux in the cluster.

(2) As gas cools in the cluster, it should be possible to identify multiphase gas including components with temperature well below the mean plasma temperature. A low-temperature component would however emit efficiently in low-ionization lines. Surprisingly, these are not evident in the X-ray spectrum of clusters (Peterson et al. 2003), limiting the maximum cooling rate to less than 10 $M_{\odot}$year^{-1} in typical systems.

Many clusters are observed to have a central dip in temperature. Although this is often assumed to be a result of high cooling rates, the temperature of the plasma more accurately reflects the underlying gravitational potential. In gravitating systems, cooling (energy loss) does not result in a drop in temperature.

3.3.2 Resolution

The lack of evidence of a sink for the cooling material has led to the search for an energy source to offset the radiated energy. If there is sufficient energy input, the system can be held in (or close to) a steady state. It has been apparent for some time that clusters of galaxies are often associated with radio sources, but the significance of this association has only recently become clear (Binney and Tabor 1995; Churazov et al. 2001; Fabian et al. 2006). It is now widely accepted that heat input from radio galaxy activity (ultimately from the accretion of mass onto a central black hole) provides the energy source that offsets the cooling losses from the ICM.

Radio galaxies are driven by AGN activity, but the activity is usually only visible through the radio frequency emission produced by the synchrotron radiation of relativistic electrons. In contrast to Quasars and Seyfert galaxies, optical line emission from the central black hole may be weak or absent. Radio observations classify such radio galaxies into FRI or FRII morphologies depending on the relative brightness of the central source compared to the hot spots at the end of the radio lobes. Clusters usually host FRI galaxies.

The classification and study of radio galaxies is a vast subject in its own right. For our purposes, we will focus on radio galaxies as a source of energy in the cluster. It should be emphasized that the power directly measured in the radio emission is small. However, it has recently been seen that the radio jets correspond to cavities in the X-ray surface brightness. This allows the total power to be estimated directly from the PV work done in displacing the ICM. The current paradigm is that the radio jet inflates a bubble that then raises in the cluster due to its buoyancy (Churazov et al. 2001). As it rises, its buoyant energy is dissipated in the surrounding medium. Additional heating may come from weak shocks as the bubble is inflated.

Observationally, the mechanical luminosity, $L_{\rm mech}$, is given by

$$L_{\rm mech} = \frac{\beta p_{\rm icm} V_c}{t_c} \tag{6.19}$$

where $p_{\rm icm}$ is the ambient pressure of the intra-cluster medium surrounding the cavity, V_c is the cavity volume (which must be estimated from its projected size), and t_c is the age of the cavity. β is a coefficient that accounts for the total work done as the cavity is inflated and rises: values between 3 and 10 are plausible. The age of the cavity t_c must be estimated from its buoyant rise time. Current estimates of the heating rate suggest that it is comparable to the cooling rate (➋ *Fig. 6-12*), and thus heating by radio galaxies may well resolve the "cooling flow paradox."

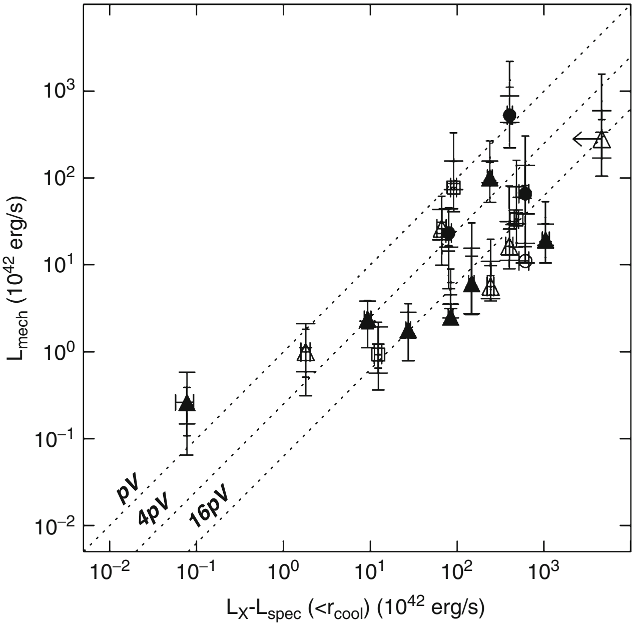


Fig. 6-12

Comparison of the heating and cooling powers of a sample of galaxy clusters (Birzan et al. 2004). Cooling luminosities are calculated from the X-ray emissivity of the cluster, while heating power is derived from cavities in the X-ray emission. *Open* and *closed symbols* differentiate systems in which the cavity is/is not filled with radio-emitting plasma. Diagonal *dashed lines* shows the total heating rate required to completely offset the cooling in each cluster (Image credit: Birzan et al. 2004)

Detailed observations show that the shutdown of star formation is not complete, and central galaxies continue to form stars at rates of typically 1–10 $M_\odot$year^{-1}, with associated emission from molecular gas (Edge 2001). The lack of a complete shutdown is unsurprising, given the intermittency of the radio outbursts and the nonspherical nature of jet emission, but the star formation observed in central galaxies makes little contribution to the growth of the stellar system.

The feedback from radio galaxies clearly plays an important role in suppressing cooling flows and thus in establishing the maximum mass of galaxies that are formed (Bower et al. 2006). Incorporating these mechanisms into theoretical models greatly improves their ability to describe the observed universe. The action of the AGN activity may also play an important role in stirring the central gas, extending its central cooling time and distributing the metals ejected by the central galaxy.

3.4 The Sunyaev–Zeldovich Effect

The Sunyaev–Zeldovich (SZ) effect arises because the hot electrons in the intra-cluster plasma scatter cosmic microwave background photons to higher temperatures. This imprints a pattern of spots on the cosmic microwave background (CMB) at the location of clusters. Depending on the frequency of observation, the spot many be more or less luminous and this makes it possible to efficiently distinguish the clusters from intrinsic CMB fluctuations and other foreground. The detection of clusters through CMB methods is currently the subject of great attention: from the ground with the South Pole Telescope and from space with the Planck mission.

The effect of the cluster is often encapsulated by the comptonization y parameter:

$$y \equiv \int \sigma_T n_e \frac{k_B T}{m_e c^2} dl \tag{6.20}$$

At low frequencies, the SZ effect cause a decrease in the CMB intensity $\delta I/I = -2y$.

One of the advantages of the SZ effect is that the decrement does not depend on redshift. Furthermore, it does not depend on the detailed clumping of the IGM, since only the pressure is important. This means that it is a powerful alternative to exploring the properties of the IGM. This will be an important avenue for future work.

4 Dark Matter

One of the most important roles of galaxy clusters has been to provide conclusive proof of the existence of dark matter. Combined with limits on the abundance of baryons from nucleosynthesis and observations of the cosmic microwave background, this implies that the dark matter is non-baryonic. We consider three methods of measuring the dark matter content of clusters below.

4.1 Galaxy Dynamics

A gravitating system of bodies obeys the virial theorem:

$$2K_e + W = 0 \tag{6.21}$$

where K_e is the kinetic energy of the system and W is the potential energy:

$$K_e = \sum_i \frac{1}{2} m_i v_i^2$$

$$W = \sum_i \sum_{j>i} \frac{G M_i M_j}{|r_i - r_j|}$$

Observations of the cluster can provide estimates of the line-of-sight velocities of galaxies. Assuming spherical symmetry, $\sigma_{3D}^2 = 3\sigma_{\text{los}}^2$. This is equivalent to a plasma temperature of

$$T \approx 7 \times 10^7 \left(\frac{\sigma_{\text{los}}}{1,000\,\text{km}\,\text{s}^{-1}} \right)^2 \text{K} \tag{6.22}$$

Similarly, the 2-dimensional distribution of the galaxies can be used to estimate the average inverse separation. For a r^{-2} density distribution,

$$\left\langle \frac{1}{r_i - r_j} \right\rangle \approx \frac{1}{0.6R}$$

where R is the virial radius of the system. Hence, the mass of the cluster can be estimated:

$$M \approx \frac{3\langle \sigma_{\text{los}}^2 \rangle R}{G} \tag{6.23}$$

More accurate calculations must allow for orbital anisotropy and a variety of other effects.

4.2 Hydrostatic Equilibrium

The X-ray-emitting plasma permits another approach to mass measurement. Assuming spherical symmetry and that pressure forces balance the gravitational forces so that the system is in hydrostatic equilibrium, we have

$$\frac{1}{\rho_{\text{gas}}(r)} \frac{dP}{dr} = -\frac{GM(<r)}{r^2} \tag{6.24}$$

where $M(r)$ is the total mass enclosed within radius r and ρ_g is the gas density.

Since the intra-cluster plasma is well described by an ideal gas with mean molecular weight μ, the gas density and temperature are related to pressure by

$$P = \frac{\rho_g kT}{\mu m_p} \tag{6.25}$$

where our assumption of hydrostatic equilibrium means that turbulence makes a negligible contribution to the total pressure. For a pure ionized hydrogen plasma, $\mu = 0.5$, and for a more realistic plasma including helium and heavier elements, $\mu = 0.6$. Eliminating pressure from the equations gives

$$M(r) = -\frac{k_B T(r) r}{\mu m_p G} \left(\frac{d \ln T}{d \ln r} + \frac{d \ln \rho_g}{d \ln r} \right) \tag{6.26}$$

Since T and ρ_g can be mapped from the X-ray data, we can see that X-ray measurements of the 3-dimesional density and temperature profiles thus allow us to determine the total enclosed mass.

For a simple example, if the cluster is isothermal and the density profile is described by $\rho_g \propto r^{-2}$ (as appropriate for (➔ 6.14) at large radius), the enclosed mass is given by

$$M = \frac{2k_B T}{G\mu m_p} R = 6.5 \times 10^{14}\, M_\odot \left(\frac{T}{10^8\,\mathrm{K}}\right)\left(\frac{R}{1\,\mathrm{Mpc}}\right) \tag{6.27}$$

More complete analyses can make use of the full temperature and density profiles. Using simulations as a guideline, a 10% correction should be included to allow for departures from hydrostatic equilibrium. The error arising from departures from spherical symmetry can also be estimated in this way. The error tends to be small because the gravitational potential is more symmetric than the underlying gas distribution. High-quality X-ray measurements thus provide a robust means of determining the mass profiles of clusters.

4.3 Gravitational Lensing

Because of the high dark matter content of clusters, they create strong sources of gravitational lensing. These result in spectacular giant arcs. However, the use of gravitational lensing for cluster is more easily applied to the weak lensing case, where the lensing effect results in the weak distortion of background galaxies. The effect on an individual galaxy is weak, but by averaging over many galaxies, a clear signal can be obtained.

The underlying principle of gravitational lensing is that the gravitational potential of the cluster gives space an effective refractive index. This deflects light rays passing through the cluster. For a circularly symmetric lens, the deflection angle, α, is given by

$$\alpha = \frac{4G}{c^2}\frac{M(<b)}{b} \tag{6.28}$$

where b is the closest distance between the light ray and the cluster and $M(<b)$ is the projected mass enclosed within this radius. We can express in terms of the geometry of the lens, $b \equiv D_L \theta_I$, where D_L is the distance to the lens and θ_I is the angular separation of the source and lens as seen by the observer.

In the weak lensing case, where $\alpha << \theta_I$, the effect is to distort the background source into a slightly elongated image. The elongation is related to the lensing potential by

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} \psi_{11} - \psi_{22} \\ 2\psi_{12} \end{pmatrix} \tag{6.29}$$

where $\psi_{ij} = \frac{\partial^2 \psi}{\partial\theta_i \partial\theta_j}$ and ψ is the lensing potential given by

$$\nabla^2 \psi = \frac{D_L D_{LS}}{D_S}\frac{8\pi G}{c^2}\Sigma$$

where $\Sigma = \int \rho dl$ is the projected mass surface density.

These equations can indeed be inverted to map the mass density of the cluster (Kaiser and Squires 1993). For a spherically symmetric cluster, we can relate the distortion to the surface density within a certain radius as

$$\gamma = \frac{\bar{\Sigma}(<b) - \Sigma(b)}{\Sigma_{\text{crit}}} \tag{6.30}$$

where Σ is the mass surface density. While the lensing signal directly measures the gravitating mass of the system, the signal is still subject to confusion by line-of-sight projections, such as the "mass-sheet degeneracy" (Bradac et al. 2004).

In the strong lensing case, the affect on the light path is more extreme and nonlinear. Several different paths through the cluster may result in local stationary points in the path length and thus correspond to images. The brightness of the different images will vary, some being brighter and others fainter. Note that the intrinsic surface brightness of these images is the same; it is their angular extent that differs, and this can give rise to large magnification factors. Although the strong lensing measurements cannot give total cluster mass (since the strong lensing effect is confined to the center of the system), they give a powerful technique for measuring the mass distribution within the Einstein radius. This can, in turn, be used to confirm the mass distribution with the radius probed by X-ray measurements.

4.4 Implications for Cosmology

The presence of dark matter in clusters is clear. All the methods are in broad agreement (see ➋ *Fig. 6-13*), with a much greater mass than can be associated with the galaxies alone, or indeed even with the diffuse intra-cluster plasma. Of course, this does not directly preclude the

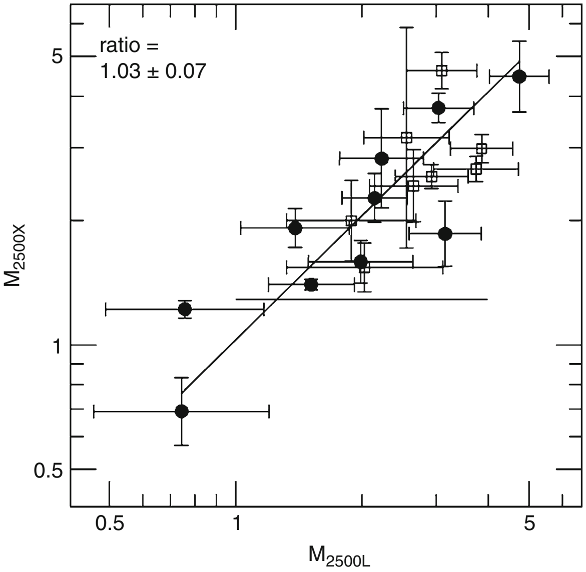


Fig. 6-13
Comparison of cluster masses derived from lensing and X-ray studies using the techniques described in the text. The figure is taken from Mahdavi et al. (2008). Masses are computed within a radius of $r_{2,500}$ and are in units of $10^{14}\ M_{\odot}$ (within this radius, the enclosed average density is 2,500 times the critical density). *Solid* and *open symbols* denote clusters with/without central temperature decrements. The *solid line* shows the relation expected if the masses are equal (Image credit: Mahdavi et al. 2008)

dark component being baryonic, but it is hard to imagine any viable way in which the required vast abundance of baryons could be hidden. For example, the necessary mass of dust would greatly exceed the limits on obscuration of background sources. We conclude that most of the mass in cluster is non-baryonic.

Moreover, observations of the microwave background and nucleosynthesis are incompatible with a significant baryonic component. Recent CMB constraints suggest $\Omega_b/\Omega_M = 0.167$ (Komatsu et al. 2011). Allowing for the small (10%) contribution from baryons locked up in stars, this is 13% higher than the observed value in galaxy clusters (Ettori et al. 2009). The remaining factor maybe due to some mass ejected from the halo. This process is clearly seen in galaxy groups, and it is thought that energy injection from AGN is responsible.

Additional cosmological constraints come from the abundance of clusters. They are a sensitive probe of the highest peaks of the primordial density field. To understand how the abundance of clusters can be used as a cosmological constraint, we must understand how clusters are formed.

5 The Formation of Clusters

A full treatment of the evolution of cosmological density perturbation is beyond the scope of this chapter. We provide a very brief outline.

One approach is to consider the evolution of the cosmological density field following the ideas of Press and Schechter (1974). This provides a simple analytical model of the most important features. In the small amplitude regime, the growth of density perturbation is linear and the different Fourier modes can be treated independently.

For a critical density universe, the perturbations, $\delta \equiv \Delta\rho/\rho$, grow as $(1+z)^{-1}$. For a more general cosmology, the growth rate is more complicated, but the principle is the same. However, as the perturbations grow in amplitude, they eventually become large enough that the nonlinear terms in the equations can no longer be ignored. At this point, the growth rate accelerates. When $\delta \sim 1$, the perturbation turns around and decouples from the cosmic expansion. It then rapidly collapses to form a virialized object.

The key idea of the Press–Schechter model is to separate the growth of the perturbation into the linear and nonlinear regimes. In this way, it is possible to count the fraction of space that collapses into objects of a given size. So that the probability that a particle is part of an object of a mass greater than M is given by

$$F \equiv p(\delta > \delta_c | R_f(M))\,[\times 2] = \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{\delta_c}{\sqrt{2}\sigma(R_f, z)}\right)\right)[\times 2] \tag{6.31}$$

δ_c is the threshold for nonlinear collapse. In the critical density of the universe, this is usually set to 1.68 (to match the behavior expected for a top hot spherical density perturbation). $R_f(M)$ is an appropriate scale corresponding to collapsed objects of mass M. An appropriate choice would be to set this to $(M/(4\pi\rho_0/3))^{1/3}$, where ρ_0 is the average density of the universe. The term $\sigma(R_f, z)$ is the rms amplitude of density fluctuations on scale R_f at redshift z. You should read the left side as giving the probability for collapse when the linear density field has been filtered on scale R_f. The growth on the density fluctuations is encapsulated in the $\sigma(R_f, z)$ term. We can simplify this expression by encapsulating the redshift and mass dependence in $\nu = \frac{\delta_c}{\sigma(R_f,z)}$.

I have included a mysterious term $[\times 2]$ factor in this expression. If we were to count only regions where the linear field exceeded the density threshold, this term would not appear. However, the formula would then only ever place half the mass of the universe in collapsed objects. Press and Schechter multiplied the expression by 2 so that all the mass is placed in collapsed objects if the threshold is sufficiently low. One argument is that we must account for regions of the universe that are close to over-dense peaks as well as the peaks themselves. As the peak collapses, it pulls in material from the surrounding region. This process does not need to be independent of scale, but the factor of 2 approximation does a remarkably good job. Another way to look at this problem is to examine the trajectories of the density of a point as a function of the filtering scale. In this way, we can avoid undercounting regions (Bond et al. 1991).

In order to derive the mass spectrum of density perturbations, we can differentiate this probability function in order to obtain the mass function of objects with mass M. If we define $f(M)$ as the commoving number of objects of mass M, so that $Mf(M)/\rho_0 = |dF/dM|$, we obtain

$$f(M) = \rho_0 M^2 \left| \frac{d \ln \sigma}{d \ln M} \right| \sqrt{\frac{2}{\pi}} \nu \exp\left(-\frac{\nu^2}{2} \right) \tag{6.32}$$

With recent advances in numerical simulations, it is possible to improve on this mass function by more carefully calibrating the terms. A detailed discussion of a better approximation to this mass function, and extension to a wider range of cosmologies can be found in (Reed et al. 2007).

Nevertheless, the simple analytic formula gives remarkable insight into the formation of structures in the universe and the growth of clusters in particular. At the present epoch, the characteristic mass of haloes is $\sim 10^{13}\, M_\odot$ and thus clusters represent extreme objects in regions of accelerated structure growth. It is possible to take this picture further and to ask how the growth of a cluster differs from that of an average region of the universe.

Of course, numerical simulations are needed to see the full 3-dimensional picture. It can be seen that clusters are knots formed at the intersection of the filaments. They grow by accreting smaller lumps (such as galaxy groups) as well as more diffuse material that drains in along the filaments. The Press–Schechter model can give insight into the growth statistics, but these can also be determined directly from the numerical models.

An interesting question to ask is, what fraction of a cluster's mass arrives in the form of galaxy groups and what fraction in mass units of individual galaxies or smaller (McGee et al. 2009). A relatively large fraction of galaxies in a cluster have previously been in a group. This raises important issues. The properties of galaxies now in rich clusters might have been largely determined by the environmental interactions at the galaxy group stage.

These theoretical developments allow us to connect together snapshots of galaxies in clusters and groups at different epochs and give us the tools needed to create an empirical history of cluster galaxy formation based on statistical arguments. Our understanding of the growth of clusters also allows us to better trace the growth of the entropy of the intra-cluster medium and to explain the nontrivial scaling of the entropy distribution with cluster mass.

The abundance of clusters is an encouraging cosmological probe. Cluster abundance should vary strongly as a function of the cosmological power spectrum amplitude. Its evolution is a strong function of the background cosmology as illustrated in ➋ *Fig. 6-14*.

The difficulty with these constraints is that cluster masses are not measured directly, but rather use proxy such as X-ray temperature, luminosity, or red galaxy content. Measurements

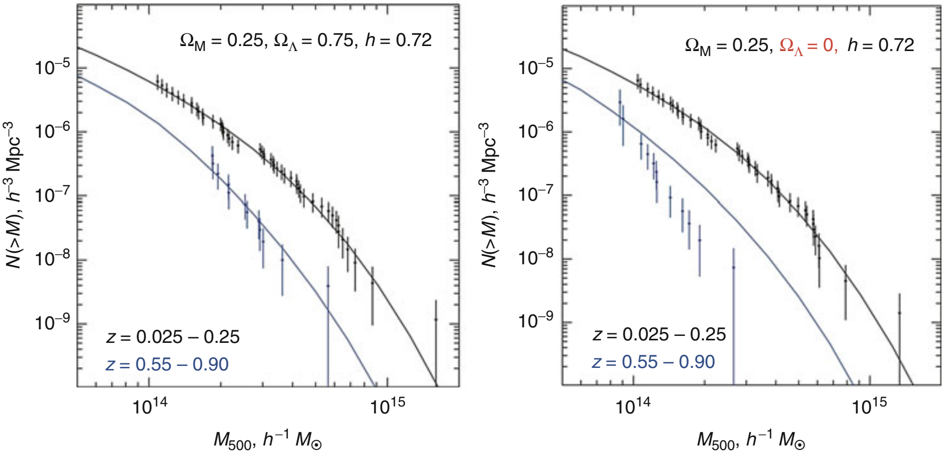


Fig. 6-14

An illustration of the cosmological constraints that may be obtained from measurements of cluster abundance, from Vikhlinin et al. (2009). The black (*blue*) lines distinguish cluster samples at high and low redshift. The two panels compare predictions (*lines*) with observational data. Note that the change of cosmology affects both the observational data (through the luminosity distance and the volume of the survey) and the model prediction (through the growth rates of dark matter haloes) (Image credit: Vikhlinin et al. 2009)

based on the SZ effect may overcome some of the biases, but to make cutting-edge constraints requires control of systematics at better than a 1% level. This may be possible in future surveys, but it is clearly challenging to improve on the tight constraints that come from observations of the cosmic microwave background. Cluster abundance constraints may have a stronger role in constraining (or measuring) non-Gaussian distortion of the power spectrum. These cause the growth rate of clusters to exceed (or lag behind) the rates estimated from the power spectrum and may provide significant insight into the nature of inflation or quintessence (Verde and Matarrese 2009).

6 Summary and the Future

6.1 Summary

This has been a brief introduction to the scientific excitement of galaxy clusters. The picture has changed considerably from the inception of cluster studies based on optical concentrations of galaxies seen in photographic plates. We now realize that the major mass component of the cluster is unseen. It has been revealed through careful dynamical measurements, the high pressure of the intra-cluster plasma in galaxy clusters, and through the distortion of background objects lensed by the gravitational mass of the system. Even amongst the baryonic component, we now realize that the stars in galaxies in the cluster make up a minor contribution. Most of the baryons in the cluster are in the diffuse, hot plasma that is trapped by the system's gravitational potential. Clusters of galaxies have given us much greater insight into the formation of the universe than just their role as a laboratory for the interactions of galaxies.

In this introduction, we have looked at:

- The optical properties and galaxies. I have emphasized the role clusters play as laboratories for the interactions of galaxies with their environment. "Galaxy ecology" is rapidly gaining momentum and understanding through the comparison of cluster (and group and field) observations over a wide range of redshifts with theoretical models. The study has much to tell us about how galaxies evolve and how star formation in the universe is suppressed.
- X-ray emission and the intra-cluster medium. Clusters of galaxies offer unique insight, because this material cannot be observed in lower mass systems such as the haloes of individual galaxies. Although a "warm–hot intergalactic medium" is needed to reconcile the observed abundance of galaxies and the cosmic baryon density, the WHIM cannot be observed directly, because it is too cold for observation with standard techniques. The intra-cluster medium offers us the best insight into the thermal history of most of the vast majority of baryons in the universe.
- Clusters of galaxies in their cosmological context. Clusters of galaxies also play an important role in cosmology, allowing us to independently confirm the cosmological parameters extracted from analysis of the cosmic microwave background. As the field progresses, the two approaches may take on complementary roles allowing a deep study of the nature of the initial fluctuations that seed galaxy formation in the universe.

Clusters of galaxies still offer a very promising avenue for these research goals and will continue to be the target of observing campaigns for many years to come. It will be exciting to see the new twists that a new generation of astronomers brings to cluster research.

References

Balogh, M. L., Baldry, I. K., Nichol, R., Miller, C., Bower, R., & Glazebrook, K. 2004, ApJ, 615, L101

Balogh, M. L., et al. 2011, MNRAS, 412, 2303

Bamford, S. P., et al. 2009, MNRAS, 393, 1324

Binney, J., & Tabor, G., 1995, MNRAS, 276, 663

Bîrzan, L., Rafferty, D. A., McNamara, B. R., Wise, M. W., & Nulsen, P. E. J. 2004, ApJ, 607, 800

Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., & Lacey, C. G. 2006, MNRAS, 370, 645

Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2008, MNRAS, 383, 93

Bradac, M., Lombardi, M., & Schneider, P. 2004, A&A, 424, 13

Brinchmann, J., Charlot, S., White, S. D. M., Tremonti, C., Kauffmann, G., Heckman, T., & Brinkmann, J. 2004, MNRAS, 351, 1151

Butcher, H., & Oemler, A., Jr. 1984, ApJ, 285, 426

Chandrasekhar, S. 1943, ApJ, 97, 255

Churazov, E., Brüggen, M., Kaiser, C. R., Böhringer, H., & Forman, W., 2001, ApJ, 554, 261

Couch, W. J., & Sharples, R. M. 1987, MNRAS, 229, 423

Covington, M., Dekel, A., Cox, T. J., Jonsson, P., & Primack, J. R. 2008, MNRAS, 384, 94

De Lucia, G., et al. 2004, ApJ, 610, L77

Dekel, A., & Birnboim, Y. 2006, MNRAS, 368, 2

Dressler, A. 1980, ApJ, 236, 351

Dressler, A., et al. 1997, ApJ, 490, 577

Edge, A. C. 2001, MNRAS, 328, 762

Ettori, S., Morandi, A., Tozzi, P., Balestra, I., Borgani, S., Rosati, P., Lovisari, L., & Terenziani, F. 2009, A&A, 501, 61

Fabian, A. C. 1994, ARA&A, 32, 277

Fabian, A. C., Sanders, J. S., Taylor, G. B., Allen, S. W., Crawford, C. S., Johnstone, R. M., & Iwasawa, K. 2006, MNRAS, 366, 417

Geach, J. E., Smail, I., Moran, S. M., Treu, T., & Ellis, R. S. 2009, ApJ, 691, 783

Gladders, M. D., & Yee, H. K. C. 2000, AJ, 120, 2148

Gonzalez, A. H., Zabludoff, A. I., & Zaritsky, D. 2005, ApJ, 618, 195

Gunn, J. E., & Gott, J. R., III 1972, ApJ, 176, 1

Haines, C. P., et al. 2009, ApJ, 704, 126

Hammer, D., et al. 2010, ApJS, 191, 143

Holden, B. P., et al. 2005, ApJ, 620, L83

Jorgensen, I., Franx, M., & Kjaergaard, P. 1996, MNRAS, 280, 167

Kaiser, N., & Squires, G. 1993, ApJ, 404, 441

Kennicutt, R. C., Jr. 1998, ARA&A, 36, 189

Komatsu, E., et al. 2011, ApJS, 192, 18

Leccardi, A., & Molendi, S., 2008, A&A, 487, 461

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. 2008, Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389, 1179–1189.

Mahdavi, A., Hoekstra, H., Babul, A., & Henry, J. P. 2008, MNRAS, 384, 1567

Maier, C., et al. 2009, ApJ, 694, 1099

McCarthy, I. G., Frenk, C. S., Font, A. S., Lacey, C. G., Bower, R. G., Mitchell, N. L., Balogh, M. L., & Theuns, T. 2008, MNRAS, 383, 593

McCarthy, I. G., Schaye, J., Bower, R. G., Ponman, T. J., Booth, C. M., Vecchia, C. D., & Springel, V. 2011, MNRAS, 412, 1965

McGee, S. L., & Balogh, M. L. 2010, MNRAS, 403, L79

McGee, S. L., Balogh, M. L., Bower, R. G., Font, A. S., & McCarthy, I. G., 2009, MNRAS, 400, 937

McNamara, B. R., Nulsen, P. E. J., Wise, M. W., Rafferty, D. A., Carilli, C., Sarazin, C. L., & Blanton, E. L. 2005, Nature, 433, 45

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613

Nelan, J. E., Smith, R. J., Hudson, M. J., Wegner, G. A., Lucey, J. R., Moore, S. A. W., Quinney, S. J., & Suntzeff, N. B. 2005, ApJ, 632, 137

Peacock, J. A. 1999, in Cosmological Physics, ed. J. A. Peacock (Cambridge, UK: Cambridge University Press), 704. ISBN 052141072X

Peng, Y.-j., et al. 2010, ApJ, 721, 193

Peres, C. B., Fabian, A. C., Edge, A. C., Allen, S. W., Johnstone, R. M., & White, D. A. 1998, MNRAS, 298, 416

Peterson, J. R., Kahn, S. M., Paerels, F. B. S., Kaastra, J. S., Tamura, T., Bleeker, J. A. M., Ferrigno, C., & Jernigan, J. G. 2003, ApJ, 590, 207

Ponman, T. J., Cannon, D. B., & Navarro, J. F. 1999, Nature, 397, 135

Pratt, G. W., et al. 2010, A&A, 511, A85

Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

Quilis, V., Bower, R. G., & Balogh, M. L. 2001, MNRAS, 328, 1091

Reed, D. S., Bower, R., Frenk, C. S., Jenkins, A., & Theuns, T. 2007, MNRAS, 374, 2

Richstone, D. O. 1975, ApJ, 200, 535

Saintonge, A., Tran, K.-V. H., & Holden, B. P. 2008, ApJ, 685, L113

Sanderson, A. J. R., Ponman, T. J., & O'Sullivan, E. 2006, MNRAS, 372, 1496

Sarazin, C. L. 1988, X-ray Emissions from Clusters of Galaxies, Cambridge Astrophysics Series (Cambridge: Cambridge University Press)

Simard, L., et al. 2002, ApJS, 142, 1

Stott, J. P., Smail, I., Edge, A. C., Ebeling, H., Smith, G. P., Kneib, J.-P., & Pimbblet, K. A. 2007, ApJ, 661, 95

Sutherland, R. S., & Dopita, M. A. 1995, ApJ, 439, 381

Verde, L., & Matarrese, S. 2009, ApJ, 706, L91

Vikhlinin, A., Kravtsov, A., Forman, W., Jones, C., Markevitch, M., Murray, S. S., & Van Speybroeck, L. 2006, ApJ, 640, 691

Vikhlinin, A., et al. 2009, ApJ, 692, 1060

White, S. D. M., & Frenk, C. S. 1991, ApJ, 379, 52

Wiersma, R. P. C., Schaye, J., & Smith, B. D. 2009, MNRAS, 393, 99

Worthey, G. 1994, ApJS, 95, 107

7 Active Galactic Nuclei

Eric S. Perlman
Department of Physics and Space Sciences, Florida Institute of Technology, Melbourne, FL, USA

1 *A Little History* 306

2 *Fundamental Properties of AGN* 308
2.1 Overall Continuum Shape 308
2.2 Emission and Absorption Lines in the Optical – UV 310
2.3 Radio Emission and the Radio-Loud/Radio-Quiet Divide 312
2.3.1 Radio-Loud Versus Radio-Quiet: A True Bimodality? 314
2.3.2 Radio Morphology 314
2.4 Infrared Emission 318
2.5 X-Ray and Higher-Energy Emission 320
2.6 Variability 322

3 *The Overall Structure: Unified Models* 325
3.1 The Basic Scheme 326
3.2 Evidence for Type 1/Type 2 Unification 329
3.3 Blazars, Jets, and Unified Schemes for Radio-Loud AGN 330

4 *Nuclear Black Holes and Accretion* 335
4.1 Bondi Accretion 339
4.2 Disk Accretion 340
4.3 Disk Coronae and the 6.4 keV Iron Line 342

5 *Emission Line Regions* 347
5.1 Reverberation Mapping 348
5.2 Physical Conditions in the BLR 350
5.3 Physical Conditions in the Narrow-Line Region 352
5.4 Photoionization 353

6 *Nuclear Obscuration: Tori, Broad, and Narrow Absorption Lines* 356
6.1 The Torus 356
6.2 Line Absorption 358

7 *Jets and Lobes in Radio Loud AGN* 366
7.1 Acceleration of Jets 366
7.2 Propagation and Dynamics of Jets at Larger Scales 370
7.3 Emission Mechanisms in Jets 373

8 *Final Remarks* 379

***References* 381**

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_7, © Springer Science+Business Media Dordrecht 2013

Abstract: Active galactic nuclei (AGN) represent an extreme stage in the life cycle of a galaxy. For a relatively short period of time (~10^8 years or less), a region less than a parsec across at the center of a galaxy produces tremendous amounts of energy, often outshining the rest of the galaxy by orders of magnitude. The most luminous of these objects are the most powerful, continuously emitting sources in the universe. The observational characteristics of AGN are reviewed, as well as how these properties are used to sort them into different classes. The evidence that supports the current paradigm, under which the central source powering AGN is a supermassive black hole, is discussed. While these are found in virtually all bright galaxies, in AGN the black holes are actively accreting matter, most likely as a result of an increased matter density in their immediate environs. A variety of accretion models are considered, particularly disk and Bondi accretion, along with the mechanisms by which material is carried inward and angular momentum is dissipated. Photoionization models for the broad and narrow emission line regions are reviewed in detail, including the physical conditions that prevail in each and their location relative to the central black hole. The evidence for unified models is presented. Under unified models, different types of AGN are related by means of viewing angle. A key aspect of these models is the presence of large-scale obscuring regions within the active nucleus. The phenomenology of nuclear absorption is discussed, both within the torus as well as in lines, along with present models for these obscuring regions. Also discussed are generation mechanisms for and the physics of relativistic jets, which are present in 10–20% of AGN. Finally, the entire picture is brought together by discussing the evolution of AGN.

Keywords: Accretion, Active galactic nuclei (AGN), BAL quasars, Blazars, Bremsstrahlung, Continuum emission, Coronal processes, Emission line regions, Galaxy evolution, Gamma-ray emission, General relativistic effects, Infrared emission, Inverse-Compton radiation, Jets, LINERs, Nuclear absorption, Nuclear structure, Outflows, Photoionization, Polarization, Quasars, Radio emission, Radio lobes, Reverberation mapping, Seyfert galaxies, Special relativistic effects, Superluminal motion, Supermassive black holes, Synchrotron radiation, Torus, Unified AGN models (Unified Schemes), Variability, Winds, X-ray emission

1 A Little History

It has been more than 100 years since the first evidence of nuclear activity in galaxies was first discovered. Interestingly, this history proceeded parallel to – rather than after – the discovery of the spiral nebulae themselves, and as with so many in astronomy, was at least partly the result of serendipity. As pointed out by Osterbrock (1999) and Shields (1999), the first documented observation of what today is called an active galactic nucleus was made by E. A. Fath in 1907 using the Lick 36-inch reflector telescope (Fath 1909). Fath was using a prism spectrograph that he had built; as part of his thesis project, he obtained low-resolution spectra of seven spiral nebulae, including M31. These observations required a titanic effort – his spectrum of M31 required 18 h of exposure. Most of his spectra were dominated by the same absorption lines seen in stars. However, one spectrum stood out: that of NGC 1068, which showed emission lines, more typical of planetary nebulae – what one would recognize today as Hβ; [O III] $\lambda\lambda$ 4363, 4959, 5007; [OII] λ 3727; and [Ne III] λ 3869. During the next two decades, several different observers – notably Pease (1915), Moore (1915), Slipher (1917), and Hubble (1926) – also noticed emission lines in spectra of NGC 1068, NGC 4151 (see, e.g., ➔ *Fig. 7-1*), and NGC 4051. Another nine such objects were known by the time that Carl K. Seyfert, then a National Research

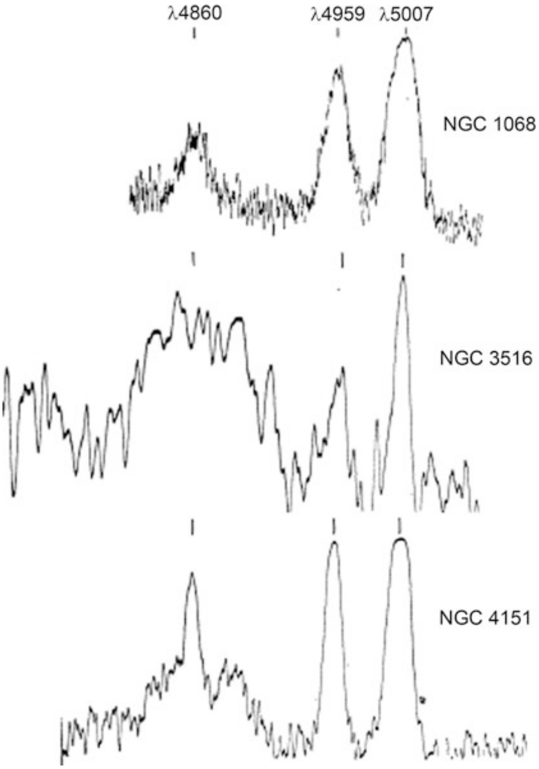


Fig. 7-1
A section of three of the early spectra of active galaxies, obtained using the Mt. Wilson 1.5 and 2.5 m telescopes by Carl Seyfert (1943, his Fig. 1). Note the bright Hβ and [O III] emission lines seen in all three spectra

Council Postdoctoral Fellow, systematized the properties of what are now called Seyfert galaxies (Seyfert 1943). As Seyfert wrote, the "most consistent characteristic" of the class was "an exceedingly luminous stellar or semistellar nucleus which contains a relatively large percentage of the total light of the system." He also pointed out that many of the anomalous emission lines were amazingly wide – up to 10,000 km s^{-1}. *Figure 7-1* shows Seyfert's spectra for three of the earliest known AGN in the region around Hβ and [O III] $\lambda\lambda$ 4959, 5007. All three of these lines are visible in each spectrum, although their strength and width vary significantly – a factor discussed later in this chapter.

It was Seyfert's work that first put forth the hypothesis that there was a distinct class of galaxies whose properties were dominated by nuclear emissions. At nearly the same time as Slipher, Pease, and Moore were undertaking their spectroscopic studies, Heber Curtis pointed out in 1918 that the galaxy M87 exhibited a "curious straight ray…apparently connected with the nucleus by a thin line of matter." At the time, Curtis's discovery was not connected to the spectroscopic evidence mentioned above – it took a large number of other discoveries for this to happen. The key development in this regard was the development of radio astronomy as a discipline. As is well known, the first receiver sensitive enough to receive cosmic radio emissions was built by Karl Jansky in the 1920s. Jansky, who at the time was working at Bell

Laboratories, was conducting a study of the sources of static in trans-Atlantic communications (Jansky 1932). His records showed two types of interference: thunderstorms and a persistent hiss that moved around in azimuth every 24 h and precessed through the sky seasonally. After further study, (Jansky 1933, 1935) concluded that the radiation came from the center and disk of the Milky Way galaxy. During the next decade, the first radio maps and surveys of the sky were done (Reber 1940a, b; Ryle and Smith 1948). By 1950, three radio sources had been identified with external galaxies (Bolton et al. 1949), namely, Virgo A (M87), Cygnus A, and Centaurus A (NGC 5128). Each of these objects was later identified optically as an active galaxy. This pointed out another facet of the AGN phenomenon – namely, that a substantial fraction have well-collimated outflows that extend out for many kiloparsecs from the nuclear regions.

Radio astronomy continued to play a role in defining our knowledge of active galaxies during the next two decades, as two bright radio sources, namely, 3C 273 and 3C 48, were the first two objects found to have systemic redshifts greater than 0.1 (Hazard et al. 1963; Schmidt 1963; Oke 1963; Greenstein and Matthews 1963). While initially there was some controversy about the nature of these redshifts (respectively 0.16 and 0.37), with some suggesting bizarre configurations such as galactic stars with a high density and hence a large gravitational redshift, the least objectional interpretation was that the redshifts were cosmological; however, this would require extreme optical luminosities for these objects, 10–30 times that of the brightest elliptical galaxy, and total emitted energies as high as 10^{60} ergs (Greenstein and Schmidt 1963).

2 Fundamental Properties of AGN

As can be seen by the previous discussion, the class of objects which today are called active galactic nuclei (or AGN) displays a rather complex phenomenology. These properties include not only bright emission lines but also strong continua (both thermal and nonthermal), X-ray, and radio emission. In this section, the observational properties of AGN are discussed, along with the taxonomy that has been developed through the years to describe subclasses.

2.1 Overall Continuum Shape

One of the distinguishing characteristics of AGN is a strong continuum that stretches from at least the infrared through hard X-rays, and in some cases up to gamma-ray energies. In ➋ *Fig. 7-2*, two examples of the broadband spectral energy distributions (SEDs) of AGN are shown. The SEDs of AGN are very different from those of typical galaxies: instead of displaying two main peaks (in the optical and far-IR, respectively, from starlight and cool dust emission) a much broader distribution is seen, with nearly equal power per decade from millimeter wavelengths through X-rays. For this reason, it is typical to characterize the spectra of AGN over a broad range of frequencies as a power law of the form $F_\nu \propto \nu^{-\alpha}$, where α is the spectral index. $\alpha = 0$ corresponds to a flat spectrum in F_ν, whereas $\alpha = 1$ describes a spectrum in which the luminosity is constant in every decade of frequency, similar to the situation seen in ➋ *Fig. 7-2* over the infrared to soft X-rays.

Two other features seen in ➋ *Fig. 7-2* are immediately obvious: the local maxima in the ultraviolet and mid-infrared, displayed in both SEDs. The first of these, often referred to as the "big blue bump," has a remarkably similar shape in most AGN, with a peak in the far-UV and

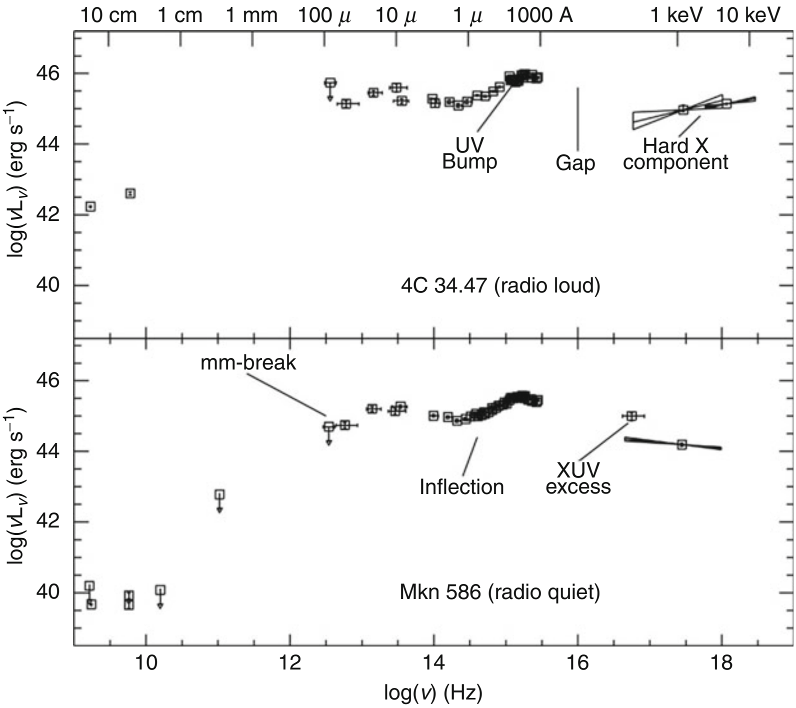


Fig. 7-2

Broadband spectral energy distributions of two typical active galaxies, one radio-loud and one radio-quiet. The two are essentially identical except in the radio, where the radio-loud object, 4C 34.47, is brighter by 3 orders of magnitude (Figure taken from Impey and Neugebauer (1988, their Fig. 1))

a typical UV spectral index of $\alpha = 0.3$ (Zheng et al. 1998; Scott et al. 2004). This translates to a U-B color of <−0.3, far bluer than that seen in "passive" galaxies and also bluer than all stellar populations except hot white dwarfs. This same characteristic also gives AGN spectra between the UV and X-rays a very characteristic shape, with spectral indices $\alpha_{\rm OX}$ clustered tightly between 1.2 and 1.8 (Tang et al. 2007). As discussed later on in this chapter, the UV bump is often interpreted in terms of emission from an accretion disk surrounding a central black hole (e.g., Shang et al. 2005 and references therein). The second "hump," believed to be due to thermal emission from warm dust at greater distances from the central engine, has a much more diverse morphology, with dust temperatures seen from tens to thousands of Kelvin, and often includes multiple components.

Emission is also seen in other bands, including a strong high-energy continuum in X-rays as well as radio emission. The strong X-ray emission is a property shared by both SEDs which, as shown later on in this chapter, is a characteristic common to all AGN, although strong obscuration in some objects can impair our ability to see this component. The radio luminosity of the two SEDs shown in *Fig. 7-2* differ radically, however, by more than 3 orders of magnitude relative to the emission seen in other bands. This points out what is known as the radio-loud/radio-quiet dichotomy, a subject that will be explored in depth later on in this chapter.

2.2 Emission and Absorption Lines in the Optical – UV

As already discussed, one of the original, distinguishing characteristics of AGN was their bright emission lines. These lines are the dominant feature of AGN spectra in the optical-UV, as shown in ➋ *Fig. 7-3*, which was generated from the Sloan Digital Sky Survey's quasar sample (Van den Berk et al. 2001). As can be seen, many bright lines are present, including the Hydrogen Lyman and Balmer series as well as lines of various ionized metal species, such as MgII, CIII, CIV, OIII, etc. Near-IR observations reveal similar lines.

The lines come in two categories: broad (FWHM > 1,000 km s^{-1}) and narrow (FWHM < 1,000 km s^{-1}). Generally speaking, it is the permitted lines (e.g., Ly α, C IV λ 1549, C III λ 1909, Mg II λ 2798, Hα, Hβ) which are observed to be broad, while forbidden lines (e.g., [O III] $\lambda\lambda$ 4959, 5007; [OII] λ 3727) are observed to be narrow – although many or most permitted lines are also observed to have a narrow core. The widths of these emission lines are generally interpreted as Doppler shifts, and thus indicative of gas motions in the producing regions. These emission lines are very different from what is normally found in external galaxies, as seen in ➋ *Fig. 7-4*, which shows the spectrum of Seyfert 1 and 2 galaxies, as well as a LINER (low-ionization emission region) galaxy, a BL Lacertae object (see ➋ Sect. 3), broad- and narrow-line radio galaxies (BLRG and NLRG, respectively), as well as a normal galaxy.

The differences between the main AGN spectral types is reasonably obvious – no broad lines are found in Seyfert 2s or NLRG, while LINERS have predominantly narrow lines, but only in low-ionization states (their lines are generally less broad as well), and the BL Lac objects

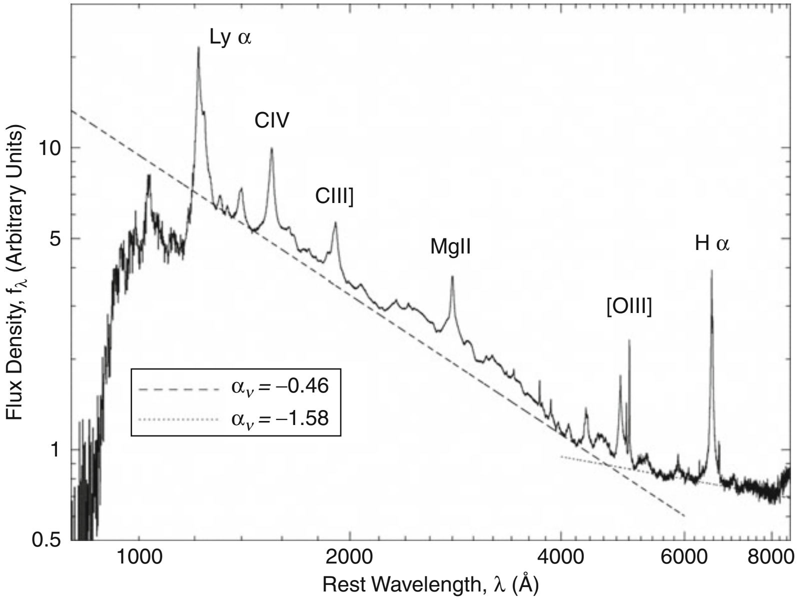


Fig. 7-3
Composite quasar spectrum, constructed from over 2,200 quasars in the SDSS DR 1. The *dotted line* indicates power law fits to the estimated continuum spectrum (From Van den Berk et al. (2001, their Fig. 3))

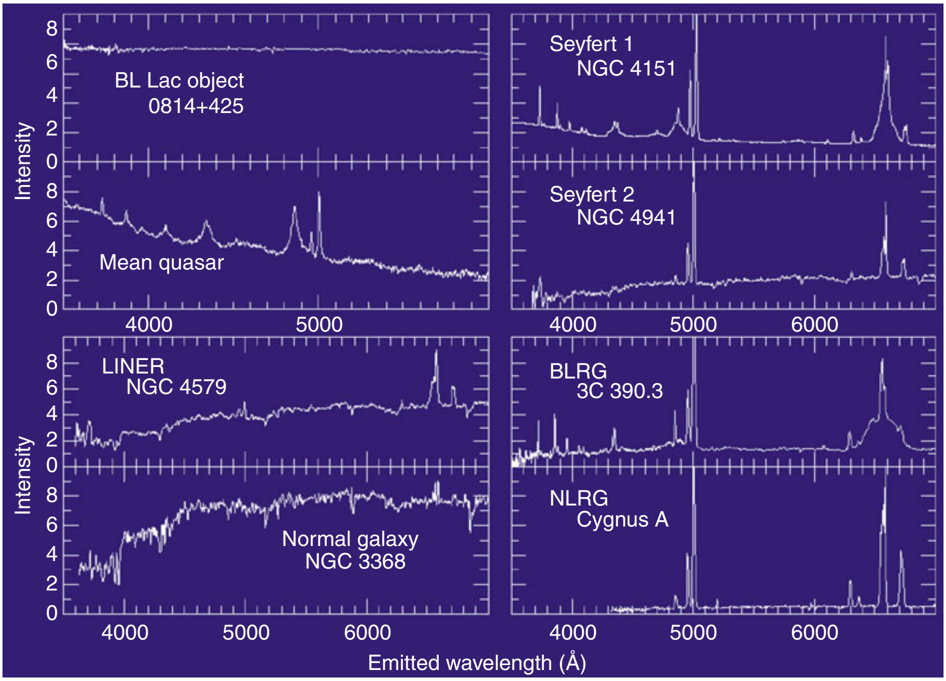


Fig. 7-4

Optical spectrum of several different types of active galaxies as compared to a normal elliptical. Note the strong emission lines present in all types except the BL Lacs and LINERs, where they are either weaker or absent. Note also that narrow-line objects completely lack the broad lines found in Seyfert 1 or BLRG-type objects (Figure taken from Keel, http://astronomy.ua.edu/keel/agn)

show a nearly complete lack of emission lines. What is not shown in *Fig. 7-4*, however, is that the same narrow lines are seen also in starburst galaxies, where they originate in HII regions surrounding hot stars. Therefore, in identifying AGN, one does need to be careful. Veilleux and Osterbrock (1987) proposed to use the line strengths of five commonly found features to distinguish between Seyferts, LINERS, and HII (Starburst) galaxies. This procedure, illustrated in *Fig. 7-5*, is quite efficient at sorting objects into AGN and nonactive objects, although there is some overlap. These "transition objects" indicate a possible link between the two phenomena.

The fact that there exists a significant population of AGN without broad emission lines was for many years a source of significant controversy in the field. Various attempts were made to explain these objects as being less luminous, but these were unsuccessful, as the bolometric luminosity of Seyfert 1 and 2 galaxies are similar. As discussed later, however, comparing the broadband spectra of Seyfert 1 and 2 objects, one sees the evidence of significant obscuration, as not only are the broad lines missing in Seyfert 2s but their optical spectra are redder, and they are also brighter in the mid-infrared (where any AGN light would be reradiated). This divide became the genesis of *unified schemes*, which will be discussed in Sect. 3.

The emission lines of AGN not only indicate the presence of a powerful ionizing and exciting source in the center, they also indicate the presence of high-velocity gas surrounding the central source. There is a strong dependency on the equivalent width of many lines on luminosity

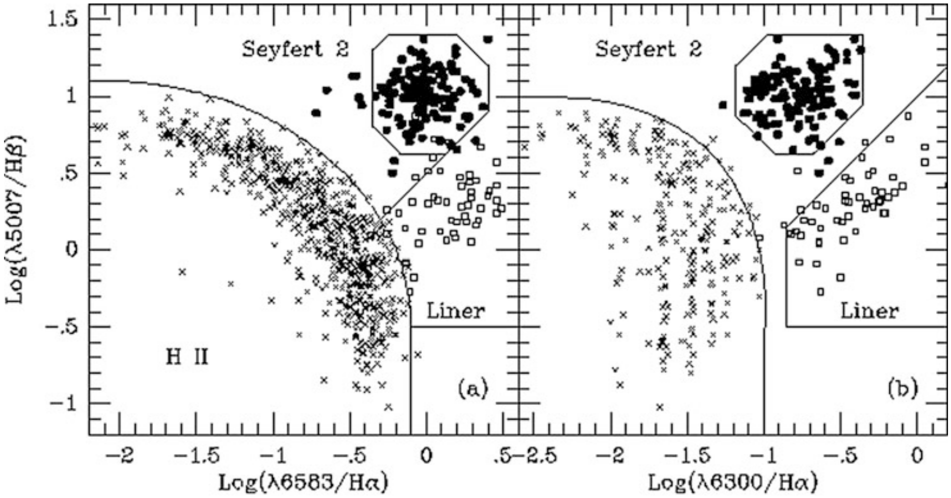


Fig. 7-5

Diagnostic diagrams for all nuclear emission line galaxies with published line ratios, excluding objects with a composite spectrum. Note that Seyfert 1, Seyfert 2, and LINER galaxies separate relatively neatly into the drawn regions, which are determined empirically. A few Seyfert 2s with weak [N II] lines appear outside the region assigned to Seyfert 2s. A few other objects falling in "intermediate" regions deserve a more detailed study (Figure taken from Veron-Cetty and Veron (2000, their Fig. 3))

(Yip et al. 2004), in the sense that more luminous objects tend to have relatively weaker lines. They can also be used to diagnose the physical state of the material in the regions producing the lines – as will be investigated in more detail later in this chapter. *Figure 7-6*, also taken from the SDSS, shows an expanded view, revealing the enormous variety of lines seen in AGN spectra. Many of the lines are blended, so to get quality diagnostics, often high resolution and signal to noise are needed. In addition to the bright emission lines, one also sees other interesting features. In particular, one sees in some of the broad emission lines a "P Cygni" type of profile – that is, the emission line is redshifted, and the blue wing exhibits strong absorption, often with a complex profile. This spectral morphology, seen in about 10% of quasars (known as "broad absorption line" or BAL quasars, discussed in Sect. 6.2), is similar to those seen in Be type stars such as their namesake, P Cygni, and in those objects they are interpreted as being due to an outflow, with the emission lines originating in the outermost shell, which is seen relatively unabsorbed, while the absorption lines originating in the portion of the wind that is approaching us. The same interpretation is invoked for AGN as well.

2.3 Radio Emission and the Radio-Loud/Radio-Quiet Divide

The SEDs of AGN come in both radio-loud and radio-quiet varieties, as shown in *Fig. 7-2*. Typical radio-loud AGN have radio continuum emission that is about 1,000 times brighter, relative to the IR-UV continuum, than a typical radio-quiet AGN. However, radio-quiet objects are significantly more numerous, amounting to 80–90% of all AGN.

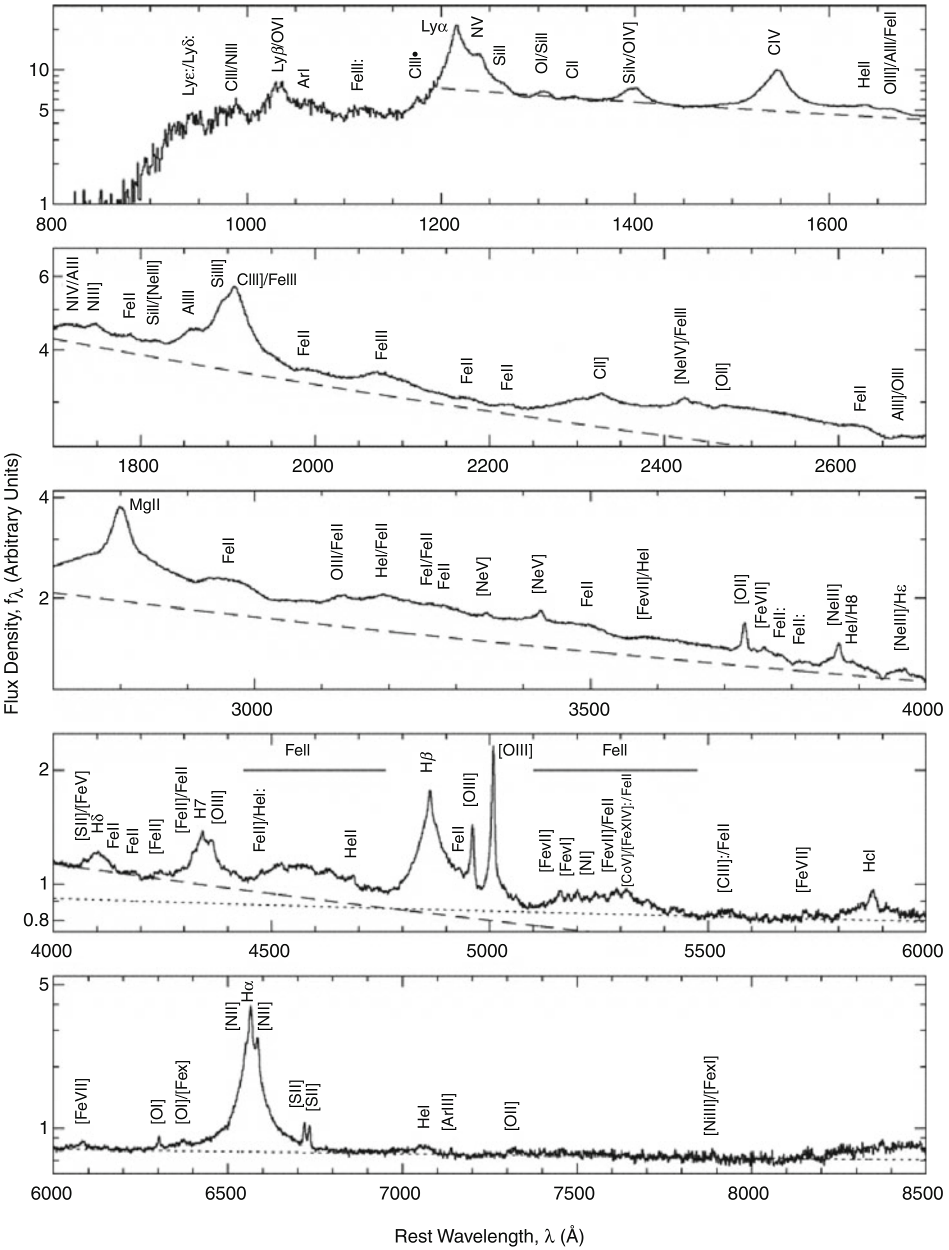


Fig. 7-6

A more detailed look at the composite quasar spectrum shown in *Fig. 7-3*. Note the wide variety of both broad- and narrow-line features (Figure taken from Van den Berk et al. (2001, their Fig. 6))

In contrast to the continuum at higher energies (far-infrared up to far-UV), which is believed to be dominated by thermal emission from various components of the AGN structure, the radio emission must have a nonthermal origin (➲ Sect. 7). The radio emission from both radio-quiet and radio-loud AGN are roughly power law in shape and display remarkably similar spectral slopes, with $\alpha \sim 0.7$ being typical of the extended structure, while $\alpha \sim 0$ is more typical of the compact, core components. The radio emission is often linearly polarized (both in the extended and core components), with degrees of polarization that can reach into the tens of percent.

2.3.1 Radio-Loud Versus Radio-Quiet: A True Bimodality?

Traditionally, the radio-loud/radio-quiet break has been drawn in terms of the ratio between the emission at 5 GHz and that in the blue (rest frame), with the dividing line being at a ratio $R \equiv F_{5\ \mathrm{GHz}}/F_{2{,}500\ \text{Å}} > 10$ (Kellermann et al. 1989; Stocke et al. 1992). The main reason why this ratio was chosen was observational – AGN were historically found by surveys in the UV or X-ray band, with radio loudness being established by correlating with a radio survey. Up until the last 10 years, the best data showed a bimodality in R, with very few objects having $R \sim 1$–10. However, more recent data are divided on this issue, and it is not certain whether this division represents a true bimodality or is a product of observational biases in the surveys (see Cirasuolo et al. 2003; Ivezic et al. 2002, 2004 for both sides of this debate, as well as ➲ *Fig. 7-7*). What is clear is that the fraction of radio-loud objects appears to depend strongly on both the optical luminosity and redshift, with the RLF at $z = 0.5$ declining from ~24.3% to 5.6% (from a sample of 4472 SDSS/FIRST quasars) as luminosity decreases from $M_{2{,}500} = -26$ to -22 and the RLF at $M_{2{,}500} = -26$ declining from 24.3% to 4.1% as redshift increases from 0.5 to 3 (Jiang et al. 2007).

It is also somewhat unclear what produces this dichotomy, real or not. Sikora et al. (2007) pointed out that there appear to be distinct radio-loud and radio-quiet sequences in the Eddington ratio versus radio-loudness plane, with both sequences showing a similar dependence of R on Eddington ratio. This led them to suggest that the main difference between the two was the spin of the central black hole, a view more recently backed up by general relativity simulations of jet formation in high and low spin environments (Tchekhovskoy et al. 2010; see also ➲ Sect. 7). An alternate view (Körding et al. 2006, 2008) is that the radio-loud and radio-quiet populations represent two different accretion states in a continuum of AGN. The idea behind this latter scenario is the behavior of X-ray binaries, which are known to switch between several states, with the same objects sometimes exhibiting radio jets and sometimes not, depending on their positions in an (X-ray) hardness/intensity diagram. Radio jets are only ejected from X-ray binaries when they are in the high-intensity and hard X-ray spectrum part of the diagram. As the timescales for phenomena around black holes should scale with the mass, one would not expect to be able to observe objects to switch from radio-loud to radio-quiet behavior. Importantly, however, the two scenarios given here may not be mutually exclusive, as pointed out by Sikora et al. (2007); it may be that the black hole spin governs the efficiency of the jet production process, but that a second factor may determine whether a jet is produced.

2.3.2 Radio Morphology

The radio morphology of active galaxies is quite varied, but several characteristics appear common to a variety of classes. Typically, one sees highly collimated outflows on one or both

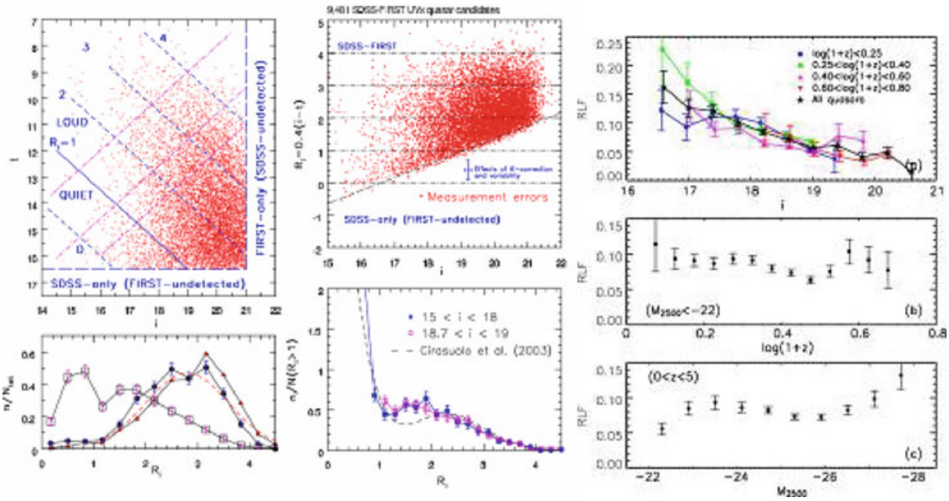


Fig. 7-7

At *left*, two analyses of the quasar radio dichotomy as carried out with a sample of 10,000 objects from the SDSS and FIRST. In the *top left panel*, which shows the source distribution in the t (radio AB magnitude) versus i (optical magnitude) plane, the diagonal *dot-dashed lines* define regions that were used to determine the R_i = 0.4 ($i-t$) distribution. The R_i histograms for these regions, marked by *filled circles* and *triangles* in the *bottom left panel*, were interpreted as evidence for a quasar radio dichotomy. The histogram marked by *open squares* shows the R_i distribution for sources with i < 18 and is shown as an example of a biased estimate of the R_i distribution. The *upper right panel* shows the R_i versus i distribution for the same SDSS-FIRST dataset as in the two *left panels* (note that this diagram is a sheared, and not simply a rotated, version of the diagram in the *top left panel*). The large dot in the *top right panel* illustrates the typical measurement uncertainty. The two histograms in the *bottom right panel* (symbols with error bars) show $p(R_i|i)$ for two ranges of i, as marked. The *dashed line* in the *bottom right panel* shows a best-fit result for $p(R_i|i)$ by Cirasuolo et al. (2003), displayed here for illustration (it is shifted left by 0.4 mag to account for different optical bands, i vs. B). Figure taken from Ivezic et al. (2002, their Fig. 1). At far right, the radio-loud/radio-quiet fraction is shown, both as a function of magnitude and as a function of redshift (Figure taken from Jiang et al. (2007, their Fig. 5))

sides of the nucleus, terminating in diffuse lobes. These lobes are often far larger than the galaxy in which the active nucleus resides and can extend for many hundreds of kiloparsecs. Several classic examples are shown in ➋ *Fig. 7-8*. 3C 341 and Cygnus A are examples of Fanaroff-Riley type II (FR II) sources: highly luminous (the classical definition is $P_{178\ \mathrm{MHz}} > 10^{25}\ \mathrm{W\ Hz}^{-1}$, as laid out by Fanaroff and Riley (1974)), with lobes that have an edge-brightened morphology, terminate in hotspots, and have fairly straight large-scale morphology. Less powerful sources ($P_{178\ \mathrm{MHz}} < 10^{25}\ \mathrm{W\ Hz}^{-1}$) do not have as extreme a morphology. Two examples of these Fanaroff-Riley type I (FR I) sources are shown in ➋ *Fig. 7-8*, namely, M87 and 3C 31. The morphological differences between FR I and FR IIs are obvious: while both classes show jets that extend for many kiloparsecs, in FR Is, one sees obvious bends and a much more edge-dimmed morphology than seen in FR IIs, usually with one or both lobes lacking obvious hotspots.

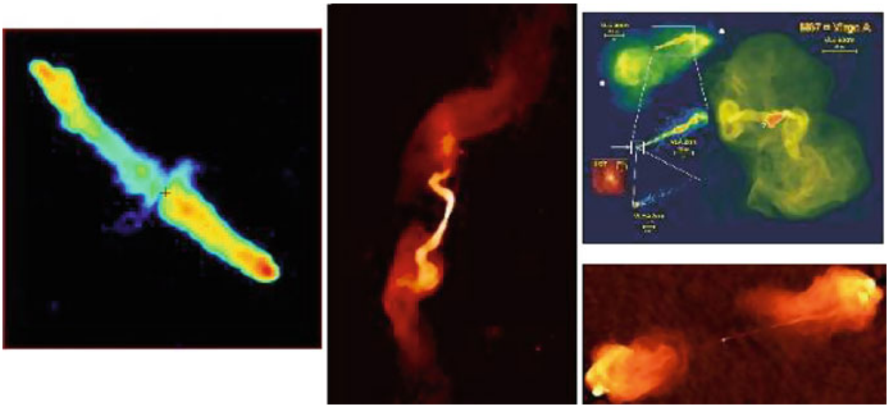


Fig. 7-8
Four radio galaxies illustrating the variety of radio morphologies seen in these objects. At *left*, we show 3C 341, an FR II radio galaxy. In the *middle panel*, we show 3C 31, an FR I radio galaxy. Note its edge-dimmed morphology and bent jets. At *top right*, we show several views of M87, an FR 1; this includes images at several resolutions from the VLA as well as an HST image. In the pseudocolor images, *red* colors indicate the highest intensity, while *blue* indicates the lowest intensity. At *bottom right*, we show the FR II galaxy Cygnus A (The three images at right are courtesy of NRAO/AUI, while the image at left is taken from Leahy and Perley (1991))

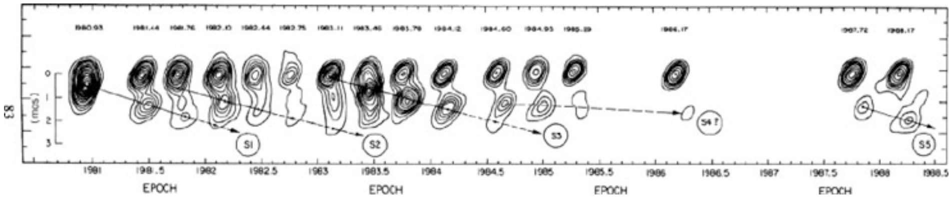


Fig. 7-9
Images of the nuclear regions of BL Lacertae over several years. Motions for several components are pointed out. Note that at the distance of BL Lacertae, 1 mas = 1.381 pc, so that an angular speed of 1 mas year^{-1} works out to an apparent speed of 4c (Figure from Mutel et al. (1990, their Fig. 1))

The jets of both FR I and FR II radio sources are collimated on small scales (within 1 pc of the central engine; see Sect. 7) and carry a major fraction of the kinetic flux and angular momentum (Sect. 4) generated by the central source with them. They affect the host galaxies of the AGN quite significantly as well, often triggering star formation in their path as well as other shocks, as witnessed by optical emission line shocks near the path of the jets in nearby radio galaxies, as well as the high-redshift "alignment effect" (see Chap. 11).

An important property of relativistic jets is that they display *apparent superluminal motion* on parsec scales. An example is shown in *Fig. 7-9* for the prototype BL Lac object, BL Lacertae. As can be seen, several components are tracked over more than 10 years; each is apparently moving at several times the speed of light.

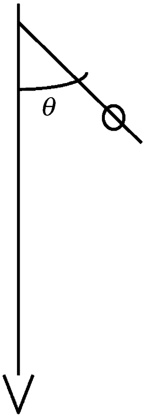


Fig. 7-10
Geometry illustrating relativistic effects in jets

This is actually a relativistic effect. The geometry is shown in *Fig. 7-10*, with an object moving at speed v at a small angle θ to the line of sight. The time of origin is chosen arbitrarily as $t = 0$. At time $t = t_e$, the component has moved a distance vt_e. As shown, the observed separation will be the transverse component of the distance,

$$\Delta r = vt_e \sin\theta \tag{7.1}$$

However at time t_e, the component has moved both in the transverse direction as well as along our line of sight and is closer to Earth than at t = 0. Photons emitted at times $t = 0$ and $t = t_e$ will reach us at times separated by

$$\Delta t = t_e - \frac{vt_e \sin\theta}{c} = t_e\,(1 - \beta\cos\theta) \tag{7.2}$$

where $\beta = v/c$ is the velocity in units of the speed of light. Thus if (7.1) and (7.2) are divided, the apparent velocity is obtained:

$$v_{\text{app}} = \frac{\Delta r}{\Delta t} = \frac{v\sin\theta}{1 - \beta\cos\theta} \tag{7.3}$$

This result was first pointed out by Sir Martin Rees in 1966 (Rees 1966) – a full decade before it was first observed. There are two results here: first of all, the observed velocity is a function both of the true velocity as well as the direction of the motion relative to our line of sight, and second, that (through a trick of geometry) it is possible to observe speeds that seem to be faster than that of light without an object ever actually traveling faster than light. Further exploration of equation (7.3) shows that a maximum velocity is found at a value $\sin\theta_{\max} = 1/\Gamma$, where γ is the Lorentz factor $\Gamma = \left(1 - v^2/c^2\right)^{-1/2}$. (It should be noted that very often in the literature one sees the capital letters as referring to bulk motion within the jet, whereas particle velocities are expressed by small letters. This convention is adopted herein). When the viewing angle $\theta = \theta_{\max}$, the apparent velocity is then $v_{\text{app}} = \Gamma v$. Thus it is quite possible for us to observe speeds much larger than c, as in M87. In fact, speeds as high as 40c have been observed in high-frequency VLBI observations of BL Lac objects (e.g., Jorstad et al. 2001, 2005).

2.4 Infrared Emission

The infrared emissions of active galaxies have been studied with a number of tools. However, the increased sensitivity of Spitzer, combined with the angular resolution possible with 10-m class ground-based telescopes, has spawned a revolution over the last decade. In the infrared, one finds much of the power emitted by AGN – typically upwards of 25%, as shown by ➤ *Fig. 7-2*. This fraction varies from 10% at a minimum to more than 50% in some objects, depending on the amount of obscuration within the AGN as well as the amount of star formation in the surrounding galaxy. Here the latter topic, which is studied in detail in the ➤ Chap. 11, is not discussed; however, it does bear mentioning in the global sense. Infrared light in AGN typically originates in warm dust, which is believed to be heated by the central AGN and also obscuring some fraction of its emission, as discussed further in ➤ Sect. 3.

The continuum emission of AGN in the infrared is characterized by three main features. These include (i) a minimum at ~1–2 μm, corresponding to the sublimation temperature of the most refractory dust (1,000–2,000 K, depending on the composition of the dust grains); (ii) an "IR bump," somewhere between a few up to 100 μm. This bump can have a variety of morphologies, from a narrow, nearly single-temperature blackbody shape in some objects to a flat morphology that requires multiple emitting components. This component is due to the thermal emission of dust, with temperatures usually between 50 and 1,000 K, and in radio-quiet objects, (iii) a steep decline ($\alpha > 3$) at large wavelengths, typical of the low energy spectrum of a gray emitter (Chini et al. 1989). This latter feature is not present in radio-loud objects, wherein the far-infrared and submillimeter nonthermal emission (discussed in ➤ Sects. 2.3 and ➤ 7) takes over.

Underlying the seemingly simple picture described above is a considerable amount of complexity. There are a broad range of spectral features in the infrared, and moreover, the mid-IR is also the site of a great deal of spatial complexity within the central regions of both the host galaxy as well as the AGN. In ➤ *Fig. 7-11*, average mid-IR spectra of Seyferts and starburst galaxies are shown, as well as six example spectra. Several things are seen in this figure. Both the Seyfert and starburst spectra possess an overarching thermal continuum, as already discussed above; however, the typical starburst possesses a considerable amount of additional cold gas, likely due to the massive amounts of star formation going on within it. Also seen in the Seyfert spectra are forbidden lines from high-ionization species, notably [Ne V] and [O IV]. This is a strong sign of a powerful exciting source in the nucleus. Also present in all of the spectra are molecular emission features, notably those due to PAH molecules. These indicate that the dusty regions that emit in the mid-IR are rich in molecular gas and also that it is not easy to completely separate the AGN from the surrounding host galaxy in this band, either physically or phenomenologically. The final set of features that are notable in all of the spectra shown in ➤ *Fig. 7-11* are a pair of features due to dusty silicates, the first located at 10 μm and the second at 18 μm. These are seen in absorption in some objects, notably Sey 2s, while in Sey 1s, they are typically seen in emission.

The stable, low-noise nature of the space environment gives space-based telescopes an enormous advantage when it comes to sensitivity. Ground-based telescopes are limited by thermal emission from the Earth's atmosphere as well as molecular features from atmospheric material, neither of which are a factor in space, where the background is orders of magnitude lower. Unfortunately, however, *Spitzer* was only a 0.85-m telescope, and as a result it does not have the angular resolution to separate the inner regions of the AGN from the nearby regions of the surrounding galaxy. To accomplish this task, one must currently resort to imaging and spectroscopy with ground-based telescopes, which are essentially diffraction limited

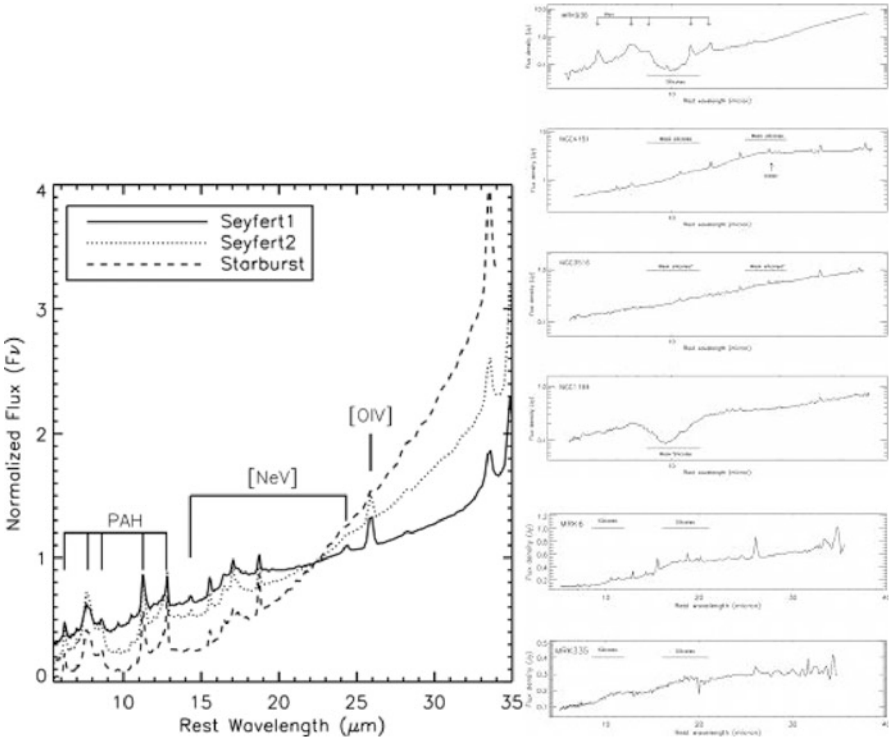


Fig. 7-11
At *left*, average 5–35-mm spectra of Seyferts and Starburst galaxies obtained with Spitzer. Figure from Wu et al. (2009, their Fig. 4). At *right*, example spectra for Seyferts in six classes, obtained with Spitzer. Figure from Buchanan et al. (2006, their Fig. 8). Spectral features are pointed out in each case

in the 10- and 20-μm windows. With ground-based 8–10-m class telescopes, angular resolutions ~0.3″ at 10 μm can be reached, a factor of 10 better than what *Spitzer* is capable of. In the mid-IR, JWST will be capable of similar resolution once launched, but much greater sensitivity; however, it will lack polarimetric capability.

An example of this work is shown in *Fig. 7-12* for the prototype Sey 2 NGC 1068 (Mason et al. 2006). The spectra cover the nucleus and central 6.0″ × 0.4″ of the ionization cones. The spectra extracted in 0.4″ (~30 pc) steps along the slit reveal striking variations in continuum slope, silicate feature profile and depth, and fine structure line fluxes, illustrating the complexity of the circumnuclear regions of this galaxy at MIR wavelengths. A comparison of photometry in various apertures reveals two distinct components: a compact (radius <15 pc), bright source within the central 0.4″ × 0.4″ and extended, lower brightness emission. The compact source is identified as the torus (see Sect. 6 for discussion) and the diffuse component as warm or hot, AGN-heated dust located mostly in the ionization cones (Sect. 4). While the torus emission dominates the flux observed in the NIR, the MIR flux measured with apertures larger than about 1″ is dominated instead by the dust emission from the ionization cones; in spite of its higher brightness, the torus contributes <30% of the 11.6-μm flux contained in the central 1.2″. Thus if one attempts to determine the torus SED using space-based data alone, one is significantly compromised by contamination from the extended emission. Mid-IR interferometry

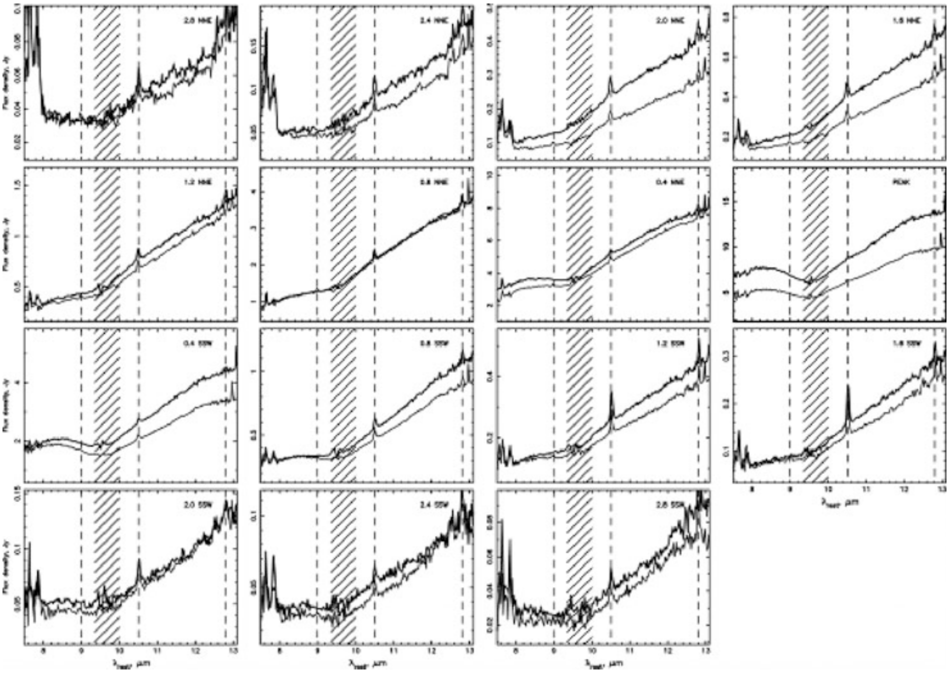


Fig. 7-12
Spatially resolved spectroscopy of the Sey 2 galaxy NGC 1068 obtained with Gemini-North + Michelle. The data are shown in 0.4″ segments along a slit that extended along the main mid-IR emission axis (PA 20°). All positions are relative to the mid-IR peak. *Black* and *gray* lines refer to two different epochs. Note how the shape of the spectrum changes within the 7.5–13-μm window – both the depth of the silicate feature (shaded region) as well as the luminosity of emission lines (e.g., [S IV] 10.5 μm and [Ar III] 8.99 μm) vary. This emphasizes how critical it is to attain high angular resolution when studying the AGN in mid-IR (Figure taken from Mason et al. (2006, their Fig. 2))

(Poncelet et al. 2006; Jaffe et al. 2004) also supports a very small central obscuring source for NGC 1068, placing a severe limit of only ~3 pc on its size. Work on a variety of other sources is showing similar results: the nuclear obscuration in AGN is concentrated in the innermost few parsecs (e.g., Radomski et al. 2003, 2008; Packham et al. 2005; Roche et al. 2006, 2007; Meisenheimer et al. 2007; Tristram et al. 2007, 2009; Poncelet et al. 2007, 2008), although in a few sources, more extended nuclear dust is seen (e.g., IC 5063, Young et al. 2007). This places severe constraints on unification models, as discussed in ➋ Sect. 6, and also directly links the regions of heaviest mid-IR obscuration to the observations of water masers by Greenhill et al. (1996) (discussed below).

2.5 X-Ray and Higher-Energy Emission

The X-ray and gamma-ray properties of active galactic nuclei are complex. Emission is seen from the galactic absorption cutoff at ~0.1 keV up to a few hundred keV. This emission contains

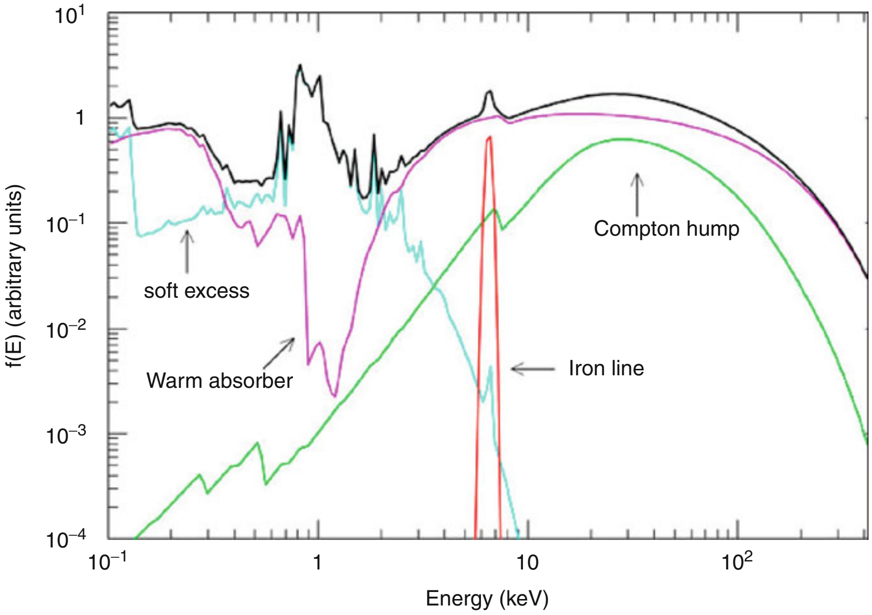


Fig. 7-13

Average total spectrum (*thick black line*) and main components (*thin grey lines*) in the X-ray spectrum of a type I AGN. The main primary continuum component is a power law with an high energy cutoff at $E \sim$ 100–300 keV, absorbed at soft energies by warm gas with $N_H \sim 10^{21}$–10^{23} cm^{-2}. A cold reflection component is also shown. The most relevant narrow feature is the iron Kα emission line at 6.4 keV. Finally, a "soft excess" is shown due to thermal emission of a Compton thin plasma with temperature $kT \sim$ 0.1–1 keV (Figure taken from Risaliti and Elvis (2004))

the signature of a number of processes, discussed in this section. *Figure 7-13* summarizes the various emission and absorption components that tend to be present in AGN X-ray spectra.

At zeroth order, the intrinsic continuum shape is a power law that extends from about 1 keV to over 100 keV. The typical spectral index of this power law is between $\alpha = 0.8$ and $\alpha = 1$, both for lower luminosity Seyfert galaxies as well as quasars. A somewhat flatter spectral shape is seen for radio-loud AGN, typically between $\alpha = 0.5$ and $\alpha = 0.7$. These spectral indices are remarkably constant with redshift, according to work done on the SDSS AGN (Green et al. 2009). The difference between radio-loud and radio-quiet X-ray spectral slopes is believed to be due to the additional inverse-Compton scattering taking place within the jet's inner regions. BeppoSAX spectra reveal that this power law cuts off exponentially at energies between 80 and 300 keV for most AGN, with the exception of some radio-loud objects (in particular blazars, which will be discussed later).

Further examination reveals that considerable complexity underlies this power law. At the soft end of the spectrum, one often observes a quasi-thermal component (the so-called soft excess), with characteristic temperature $kT \sim$ 0.2–1 keV. A second feature is seen between 10 and 50 keV, namely, a "hump" believed to be due to Compton reflection in the accretion disk's corona (see Sect. 4.3). Tied to this reflection "hump" is a warm absorber component seen in about ~50% of Seyfert 1 galaxies. This feature is quite complex, featuring a variety of line features, most notably "edges" due to hydrogen and helium-like oxygen, but also a wide variety of other highly ionized species, visible in deep, higher-resolution spectra. A variety of emission lines are also present in these objects. Most prominent among the emission lines is the Fe Kα line at a rest energy of 6.4 keV; however, a number of other emission lines are also seen in deep, higher-resolution spectra, including K- and L-shell features of iron, oxygen, silicon, magnesium and other elements. An example of a high-resolution spectrum is shown in ➲ *Fig. 7-14*.

These generalities apply to many objects; however, in a large number of AGN, one sees absorption much heavier than galactic. Obscuration in the X-rays can be due to photoelectric absorption (which dominates below a few keV) and Compton scattering (dominant above 7 keV). The observed spectral properties will depend on the amount of absorbing material present along our line of sight: column densities below $\sim 1.5 \times 10^{24}$ cm^{-2} produce a cutoff at energies between 1 and 10 keV, with higher-energy cutoffs observed in more obscured objects. In this case, the source is called "Compton thin." However, at higher column densities, the source is optically thick (and hence faint) up to several tens of keV. In this case the source is called "Compton thick," and the main spectral features seen are a prominent Fe Ka line with equivalent width ~1–3 keV and a reflected and/or scattered continuum. For less obscured sources, the Fe Kα line typically has a much lower equivalent width. These properties are summarized in ➲ *Fig. 7-15*.

2.6 Variability

Another important property of AGN is their variability, a property first discovered in the late 1960s (Cannon et al. 1968; Kinman 1968; Pacholczyk and Wyemann 1968; Barnes 1968; Selove 1969; Zaitseva and Lyutyi 1969; Babadzhanyants and Hagen-Thorn 1969) in the optical, and a couple years later in the radio (Stull 1970; Ross 1970). This variability has since been found to extend to all wavebands (see Ulrich et al. 1997 for a review), in both continuum and line emission. Variability in AGN is seen on all timescales (see ➲ *Fig. 7-16* for examples), ranging from years down to days, with a variety of amplitudes as well. In some objects (BL Lacs and blazars), violent variability is a defining characteristic, with flux varying by factors of several in short time periods.

Variability on such short timescales puts tight constraints on the size of the emitting region – since nothing can move faster than the speed of light, an object that varies within a few days must be no larger than a few light-days in diameter. This topic will be discussed in depth below, along with the nature of the central object and the need for a black hole. However, it is worthwhile to mention at this time that the variability of AGN is one of the strongest reasons behind the fact that an alternate model, advocated by Terlevich et al. (1992), is not well regarded in the literature. Under the Terlevich et al. model, AGN are viewed as giant young stellar clusters, and variability must be explained as the result of nova and supernova explosions generating rapidly evolving remnants due to the interaction between their ejecta and a high density circumstellar environment. This model is supported by the overall similarity of the optical spectra and compact supernova remnants (e.g., Filippenko 1989). What this model has difficulty explaining is not the overall character of the variations (Aretxaga and Terlevich 1994; Cristiani et al. 1996), which can be explained by a combination of complicated processes that could arise either in the structure discussed in this chapter or in separate, interstellar processes such as novae or supernovae, but rather the combination of small variability timescale and massive luminosity. For quasars, the variability properties seen require that more than 10^{12} solar luminosities, spanning

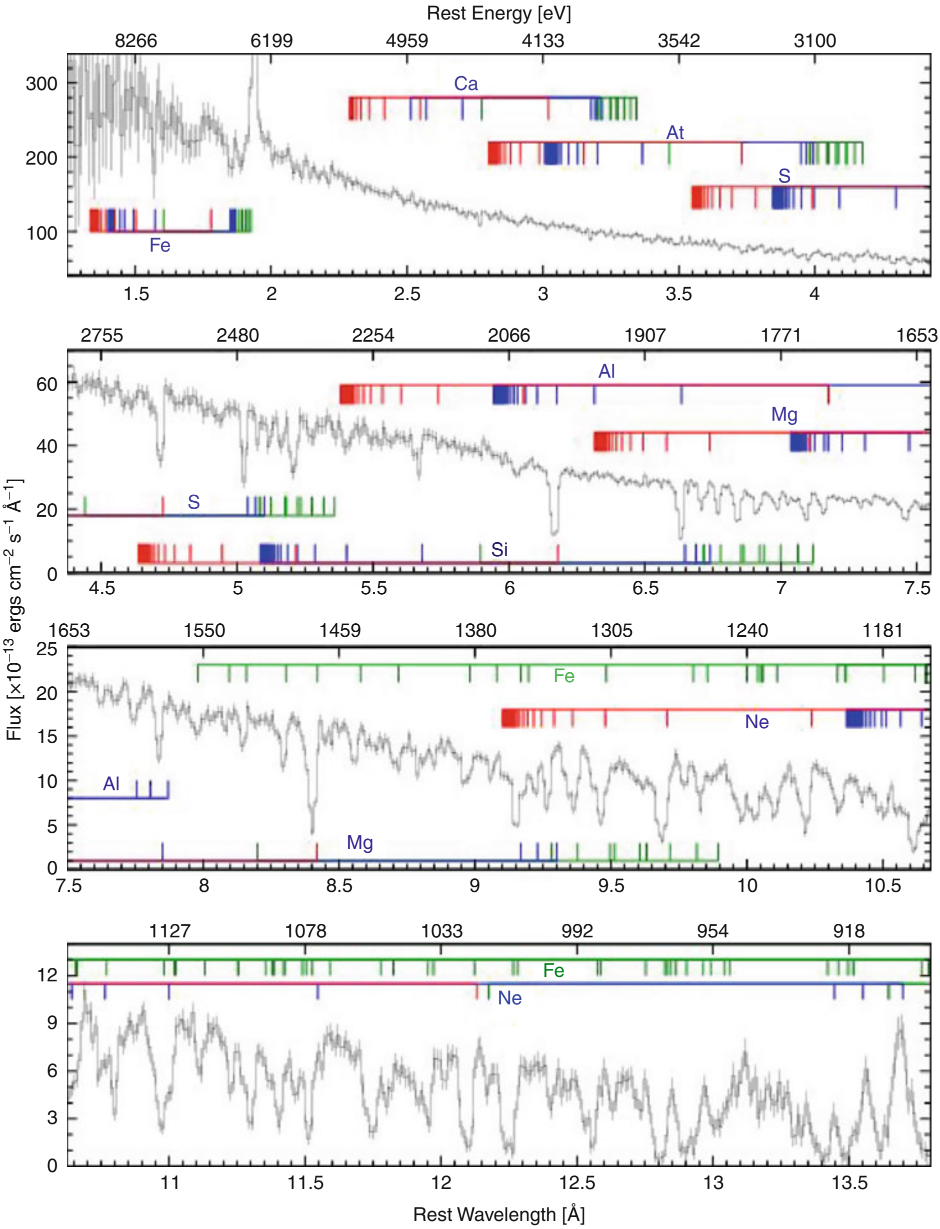


Fig. 7-14

A high-resolution X-ray spectrum of the Seyfert 2 galaxy, NGC 3783. This spectrum was obtained with the high-energy transmission grating spectrometer (HETGS) aboard the Chandra X-ray Observatory (Kaspi et al. 2002). Note the wide variety of line features

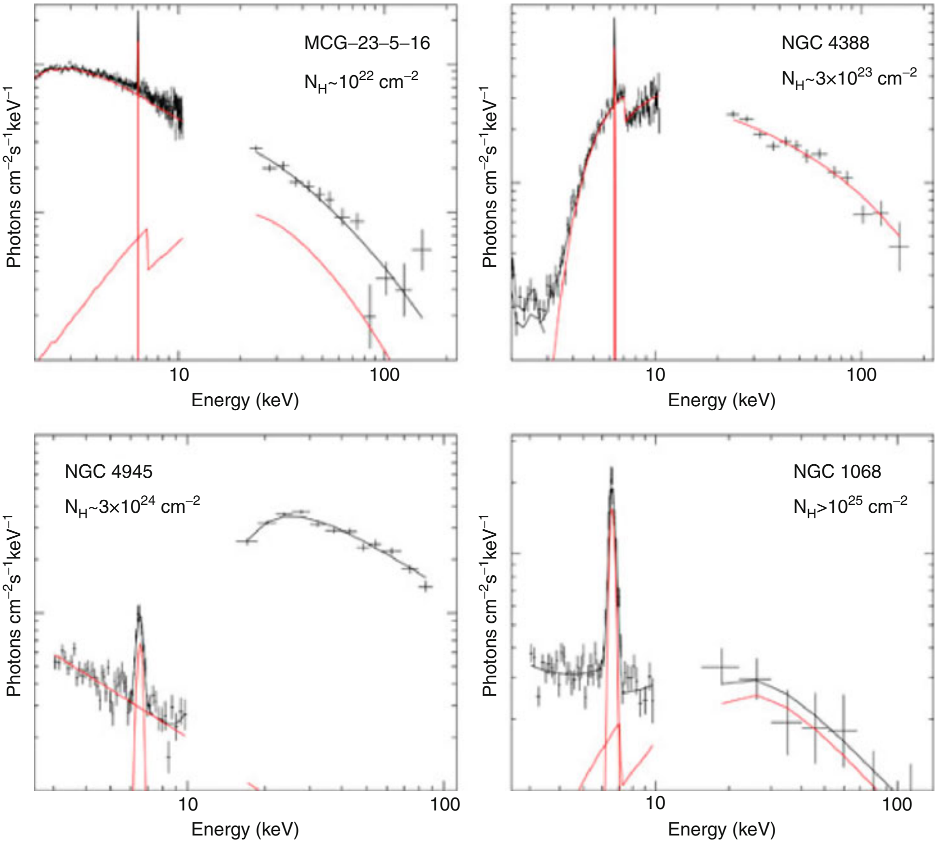


Fig. 7-15

Four 2–100-keV BeppoSAX best-fit X-ray spectra of Seyfert 2 galaxies. Main components of the best-fit models are also shown. MCG-5-23-16 and NGC 4388 (Risaliti 2002) are "Compton thin," that is, they are dominated by the primary emission down to a few keV. In MCG-5-23-16, a cold reflection component also gives a measurable contribution. The continuum in the Compton-thick source NGC 4945 (Guainazzi et al. 2000) is due to a warm reflection component in the 2–10-keV range, while at higher energies the intrinsic component emerges. Note the high ratio between the 10–100-keV and the 2–10-keV emission as compared with the Compton-thin sources. NGC 1068, also Compton-thick (Matt et al. 1997), shows a cold reflection and a warm reflection component. Equivalent widths of the iron K*α* line are ~100 eV in MCG-5-23-16, ~500 eV in NGC 4388, and 1–2 keV in NGC 4945 and NGC 1068

several decades in frequency, be radiated in a region no larger than a few light days. Explaining such large variations by a series of supernova-related explosions requires an extreme stellar density, large enough that the cluster would be unstable to gravitational collapse over short timescales.

Importantly, AGN variability is often correlated across wavebands (e.g., *Fig. 7-17*). This indicates that the processes responsible for the emission in these different bands must be connected somehow. Interestingly, however, the variability seen in different bands,

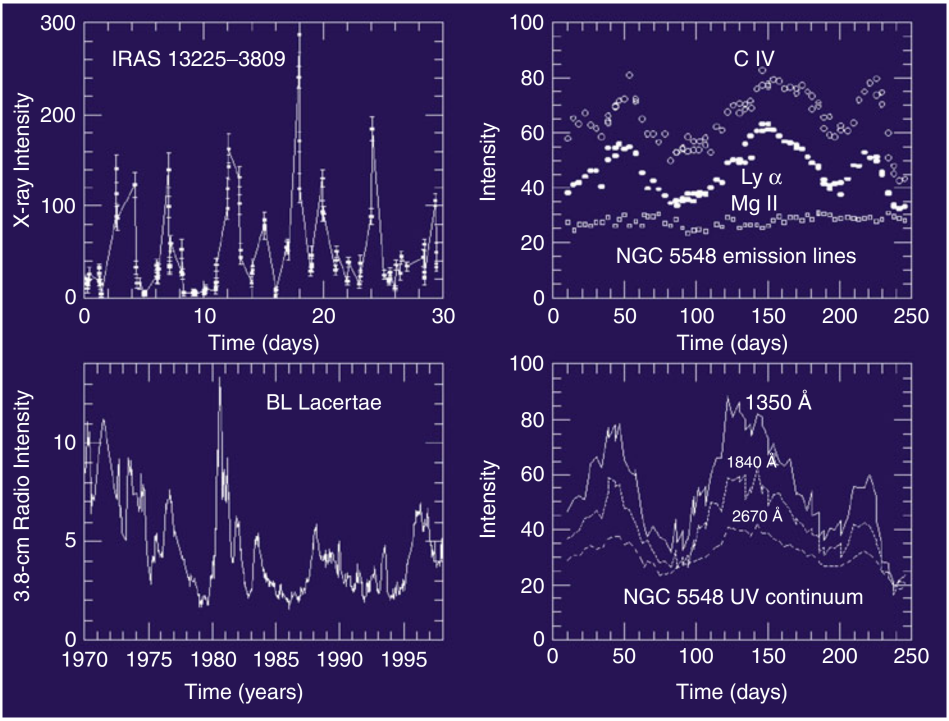


Fig. 7-16

Examples of variability in different active galaxies. The examples include two Seyferts and one BL Lacertae object (Figure taken from W. Keel http://astronomy.ua.edu/keel/agn)

while correlated, is offset somewhat in time – that is, very often it is found that shorter wavelengths lead the emission seen in softer bands.

3 The Overall Structure: Unified Models

With the large zoo of AGN properties outlined in the previous section, there is a strong motivation to explain them all through a single, unified structure that has a minimum of different features. It was out of this desire that unified schemes were born. From the outset it was clear that any unified scheme had to be able to generate the massive luminosity observed within a small region – light-days or smaller. The scheme also had to be reasonably stable for millions of years. That last fact rules out any origin in a collection of normal stars (as proposed by Terlevich in 1992), as the generation of up to 10^{12} solar luminosities by thermonuclear fusion would require of order 10^{12} solar masses within that small volume (thus being unstable to gravitational collapse in a very short time) and would have no way to explain the variability without positing the existence of novae or supernovae. The modern view centers around a supermassive black hole and posits a nonaxisymmetric system, where the properties observed depend entirely on the perspective.

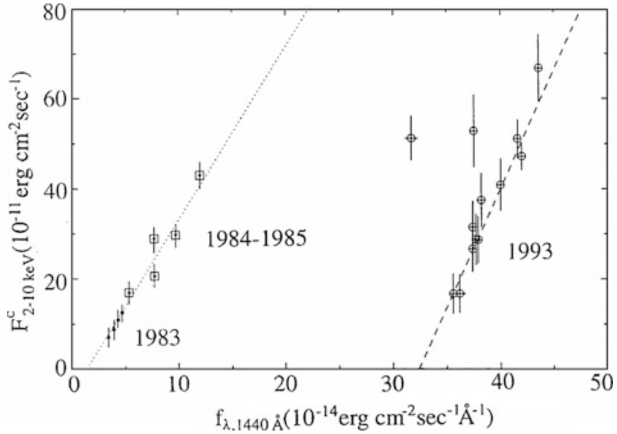


Fig. 7-17

X-ray flux (absorption corrected, 2–10 keV) versus UV flux (1,440 Å) for NGC 4151. *Right*: ASCA observations in November 30-December 13, 1993, with best-fitting linear correlation. *Left*: EXOSAT observations in December 16, 1984, to January 28, 1985 (*large crosses*), and November 7–19, 1983 (*small crosses*), with best-fitting linear correlation. The good correlation of the UV and X-ray flux on a timescale of weeks and months breaks down at long (years) and short (days) timescales (Ulrich et al. (1997), their Fig. 3)

3.1 The Basic Scheme

The basic structure, illustrated in *Fig. 7-18*, is seen in both radio-loud and radio-quiet objects and features a supermassive black hole at the center, accreting material, possibly (although not necessarily; see Sect. 4) through a disk. Accretion is an elegant way to generate enormous luminosities within the small volumes required by the observed variability as it offers the potential of converting up to half of the gravitational potential energy from accreting matter into light and heat in the process. Surrounding the accretion disk are various layers of warm gas, which would explain the broad and narrow emission line features. These two sets of gas clouds need to be located at vastly different distances from the central black hole, because, as indicated in future discussion, detailed modeling indicates that the physical conditions in the two regions are vastly different.

A critical feature of this model is the presence of a large-scale obscuration region. *Figure 7-18* illustrates this obscuration as having a uniform, optically and geometrically thick toroidal structure, which is the simplest model. However, this geometry is by no means required, and in fact the most modern data indicate that a patchy geometry may be more likely, as discussed in Sect. 6. This obscuration region would need to be dusty and possibly molecular gas-rich, but whatever its geometry, has gained the name "torus" in the literature. Its key feature is to obscure the view of the broad-line clouds from some points of view. Objects that are viewed through the dusty, obscuring clouds would be seen to have Seyfert 2 type spectra, whether with or without a radio jet, while objects viewed at a more direct angle would have broad-line spectra, more typical of Seyfert 1s. Intermediate objects would be possible depending on how much obscuring material was along our line of sight.

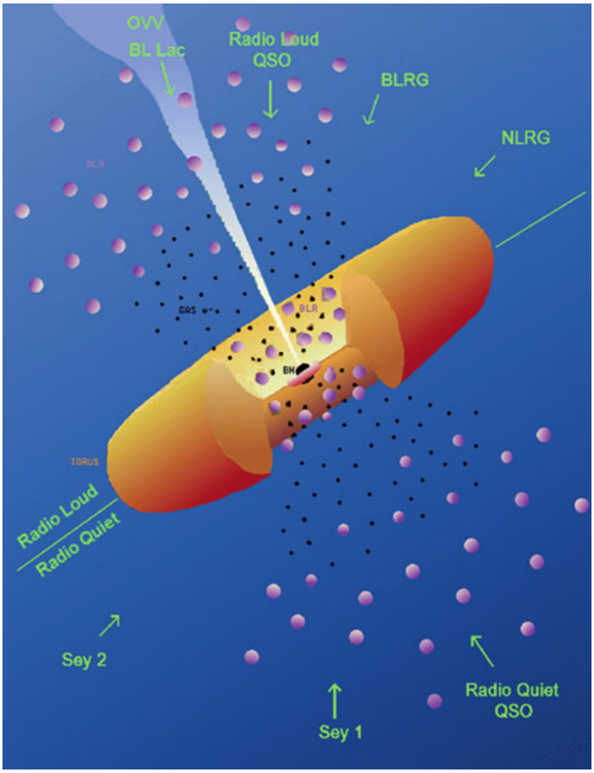


Fig. 7-18

The basic structure for active galaxies as postulated by unified schemes. This figure is not to scale, nor is the geometry for all regions settled, as discussed in the text. Nevertheless, it illustrates the basic geometry and aspect-dependent nature of the scheme (Figure courtesy of P. Padovani)

The final piece of this picture, present only for radio-loud objects, is bipolar jets that emerge from the nuclear regions at relativistic speeds. For objects where the jet was seen close to our line of sight, relativity would increase the apparent luminosity of the source so that the source was dominated by the jet properties. The result in this case would be what is known as a *blazar*, a radio-loud object with weaker or no emission lines, apparent superluminal motion and extreme variability. The most extreme of these observationally are the BL Lacertae objects, which are distinguished by their featureless optical spectra (see example in *Fig. 7-4*). Under unified schemes, BL Lacs represent low-luminosity, FR 1 radio galaxies viewed at small angles to the jet axis, while a subclass of quasars, the FSRQ or flat-spectrum radio quasars (also called OVV or optically violent variables), represent the FR 2 radio galaxies viewed at small angles to the jet axis. FSRQ are more luminous than BL Lacs and do not have their featureless optical spectra but share with BL Lacs the properties of superluminal motion and violent multiwaveband variability.

Special relativity has a number of important consequences for high-velocity jets. A little manipulation will show that a number of other interesting effects are observed for a relativistically approaching object (i.e., a jet component). The observed is also distorted for these sources, as are the frequencies of the radiation observed:

$$\Delta t_{\mathrm{app}} = \delta^{-1}\Delta t \tag{7.4}$$

$$\nu_{\mathrm{app}} = \delta\nu \tag{7.5}$$

where the commonly used *Doppler factor* is $\delta = [\Gamma(1-\beta\cos\theta)]^{-1}$. However, perhaps most importantly, the luminosity one would infer (if one naively assumes uniform illumination over 4π steradians) differs greatly from that truly emitted:

$$L_{\mathrm{app}} = \delta^{p}L \tag{7.6}$$

where the exponent p depends on the source's morphology and spectral index α. This last relation, often referred to as *Doppler boosting* or *beaming*, arises because the product of intensity and frequency cubed, $I\nu^3$, can be shown to be a Lorentz invariant. For a spherical source, $p = 3 + \alpha$, whereas for a cylindrical morphology, $p = 2 + \alpha$. Thus for a relatively modest bulk Lorentz factor $\gamma_{\mathrm{bulk}} = 10$, it is possible to infer luminosities up to 10,000 times higher if one is in the beaming cone. In ➋ *Fig. 7-19*, all three of these effects are plotted (along with

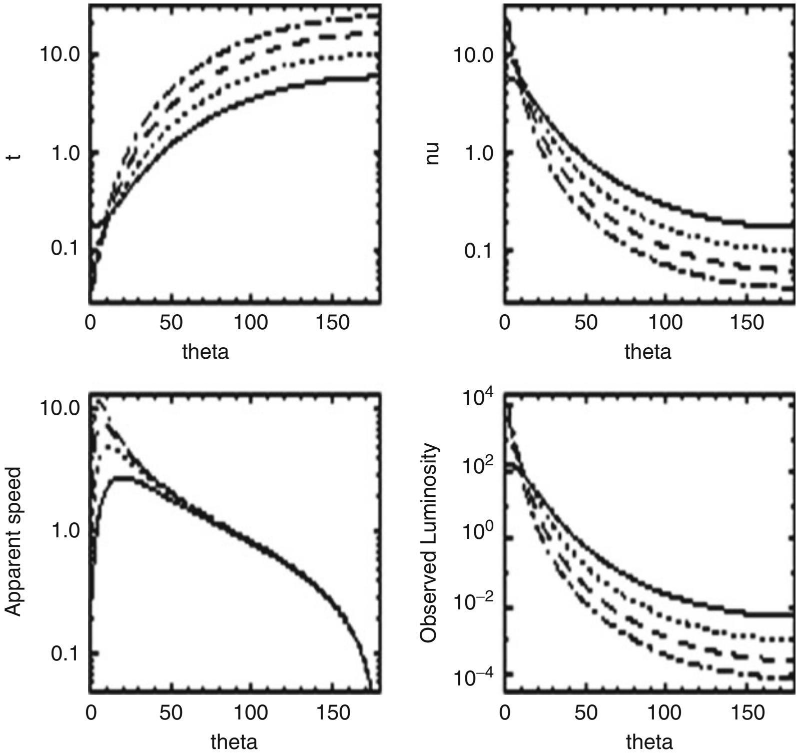


Fig. 7-19

Relativistic effects predicted by unified schemes: time dilation (*upper left*), frequency shift (*upper right*), apparent superluminal motion (*lower left*), and Doppler boosting (*lower right*) plotted versus the viewing angle in degrees. The tracks shown represent Lorentz factors $\gamma_{\mathrm{bulk}} = 3, 5, 8, 12$

superluminal motion, previously explored in ➲ Sect. 2.3) versus the viewing angle θ for $\gamma_{\text{bulk}} = 3, 5, 8, 12$. As can be seen, significant boosting is obtained for a fairly large range of viewing angles, and even for relatively modest values of γ, one can easily obtain significant Doppler boosting.

3.2 Evidence for Type 1/Type 2 Unification

The evidence supporting this view, known variously as *AGN unification* or the *unified scheme*, is strong and comes from a variety of avenues. In this section and the next, some of the strongest evidence for unified schemes is described.

If indeed the broad-line region of Sey 2 galaxies is hidden from view by dusty obscuration regions, one would expect to be able to directly detect evidence of silicates in the infrared spectra of these objects, as well as emission features from molecular features such as PAH and H_2. Observations of numerous type 2 AGN with *Spitzer* have revealed direct evidence of these features. ➲ *Figure 7-20* shows an example of such a spectrum (Armus et al. 2006), specifically for the Seyfert 2 NGC 6240. As can be seen, this spectrum shows strong silicate absorption features around 10 and 18 μm. Equally important, these same features are detected in *emission* in Seyfert 1s, exactly as would be predicted if dust clouds are obscuring our view of the central regions in these objects.

It should be noted, however, that while the Spitzer spectra just mentioned indicate the presence of an optically thick obscuring structure, they do not say anything about *what* is being obscured. Thus a second piece of evidence is required. This is provided by spectropolarimetry (Antonucci and Miller 1985; Antonucci 1983; Miller et al. 1991). Spectropolarimetry is a key diagnostic because it takes advantage of the fact that the obscuring region will scatter some fraction of the background region's light, and this scattered light will be polarized. ➲ *Figure 7-21* shows the detection of this scattered light for the Seyfert 2 galaxy NGC 1068 (Antonucci and Miller 1985). As can be seen, while the total intensity spectrum of this object shows only very weak Hβ and other Balmer lines, the broadness of these features is not obvious in total intensity. However, a completely different picture emerges in the polarized light spectrum, where broad Balmer lines typical of type 1 objects are seen, while the narrow [O III] lines are unchanged in strength. This directly indicates that the light intercepted and scattered by the dust in the torus, comes from the broad-line region.

Subsequent observations have revealed polarized broad-line emission in many other type 2 objects, both radio-loud and radio-quiet. In radio-loud objects, there is additional information, allowing us to infer the orientation of this obscuring medium relative to the radio jets. For example, in 3C 234 (Antonucci 1984) and IC 5063, scattered broad-line emission is detected, with the plane of polarization being perpendicular to the axis of the jets. This is expected if a type 1 nucleus is at least partially obscured by optically thick dust clouds whose axis coincides with the radio jet axis. Interestingly, spectroscopy of narrow-line radio galaxies also reveals that the luminosity of the obscured region is quasar-like (e.g., di Serego Alighieri et al. 1994; Goodrich and Cohen 1992; Antonucci et al. 1994, 1996), although not all NLRG have polarized optical broad lines (in particular Cygnus A, Jackson and Tadhunter 1993, although Antonucci et al. 1994 suggested that this may be due to dilution by the local optical continuum at the scattering site).

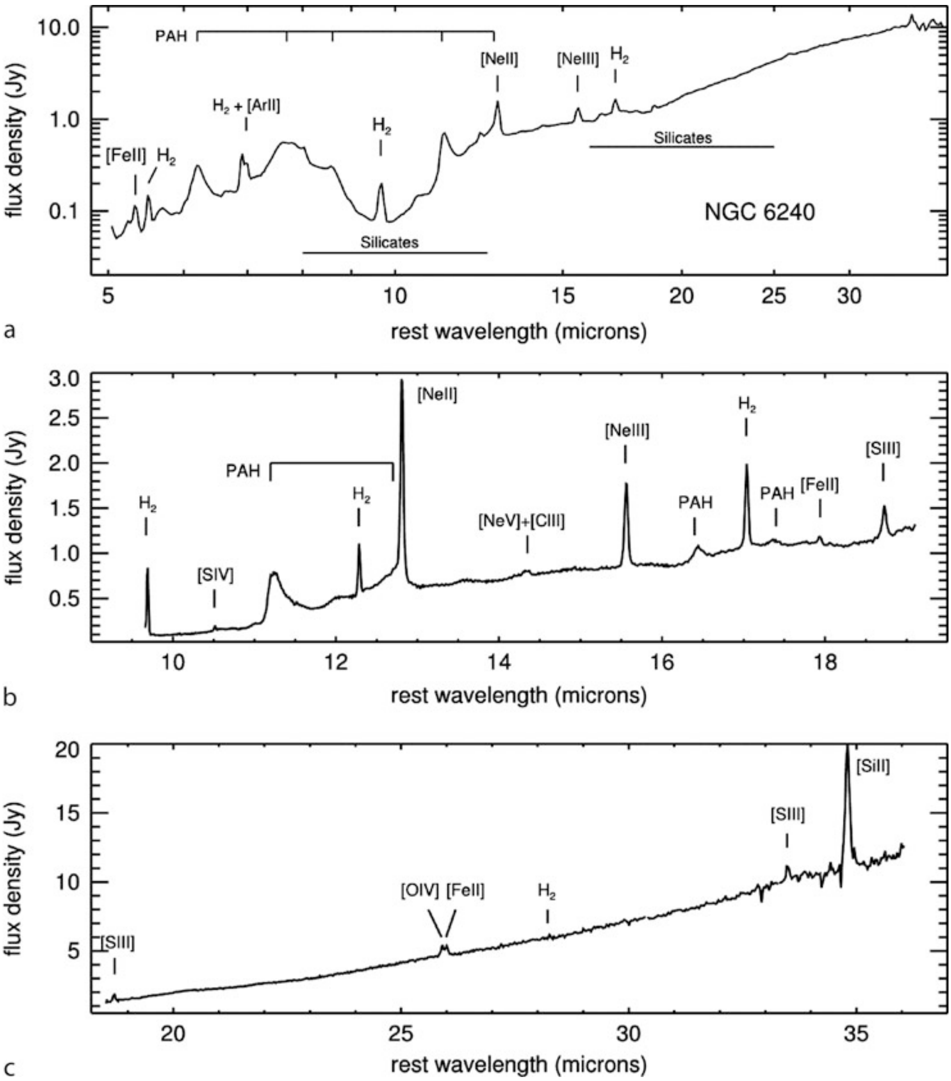


Fig. 7-20
Spitzer spectrum of the Seyfert 2 galaxy NGC 6240, revealing silicate absorption features as well as emission features from molecular gas. At *top*, the Short-Low spectrum is shown, while the *middle* and *bottom* panels show the Short-High and Long-Low spectra, respectively (From Armus et al. (2006, their Fig. 1))

3.3 Blazars, Jets, and Unified Schemes for Radio-Loud AGN

As has been seen, when an emitting plasma is moving at speeds close to that of light, relative to a fixed observer, a number of special relativistic effects are observed. Among these are superluminal motion as well as relativistic beaming. Therefore an important element in the

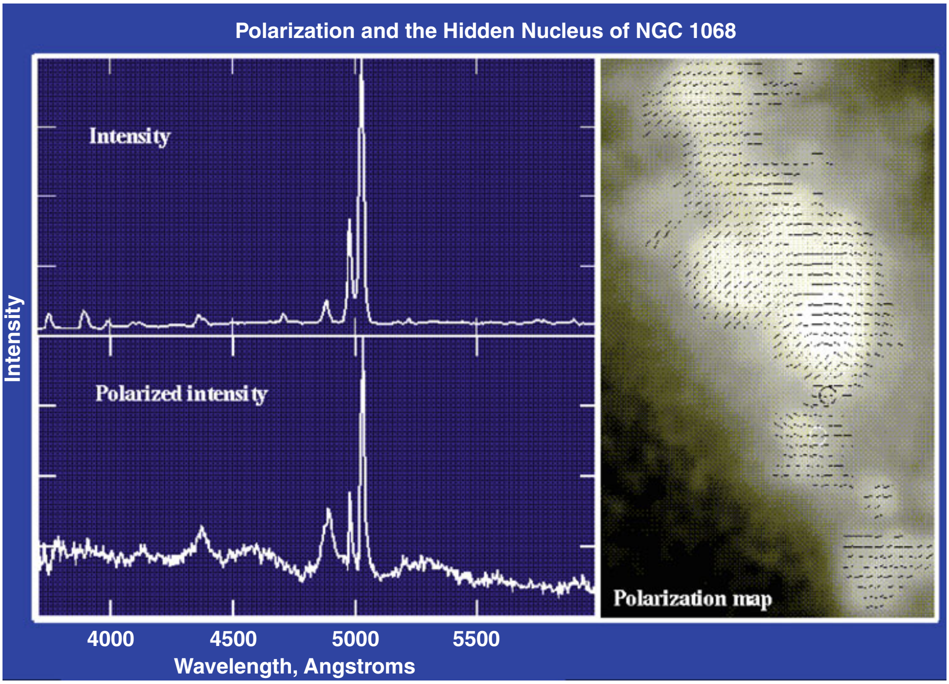


Fig. 7-21
Polarized broad-line and continuum emission from NGC 1068. As shown, broad Hβ emission is unambiguously seen in the polarized light spectrum but not in total intensity. The *right-hand panel* shows the polarization map from imaging polarimetry (Figure taken from W. Keel, using data from Antonucci and Miller)

verification of unified schemes is to see whether radio galaxies contain evidence of relativistic phenomena. This can be done in two ways: both by searching directly for those phenomena as well as by looking at the luminosity functions of both radio galaxies (known as the parent population under unified schemes) and blazars to see if the introduction of a simple relativistic beaming term can reproduce the features. In addition, since relativistic beaming is thought to explain the rapid variations and high polarizations observed in blazars and is also required to explain their GeV and TeV emission (Maraschi et al. 1992); these properties must also be present in unbeamed radio galaxies, albeit at lower levels.

Superluminal motion and violent variability have been observed in the radio galaxy M87, an FR 1 which is believed to be a misdirected BL Lacertae object (Tsvetanov et al. 1998; Harris et al. 2003). *Figure 7-22* shows two important results from multiyear monitoring of M87 with the *HST, Chandra* and *VLA*. Several regions of the M87 jet show apparent superluminal motion, most prominently in HST-1, a knot located 60 pc (projected distance) from the core. As shown in *Fig. 7-22* (left-hand panels), in this region four optical knots have been observed with apparent motions at speeds of ~6c (Biretta et al. 1999). Somewhat slower speeds have been found in the radio for this region, as high as 4.3c (Cheung et al. 2007; Biretta et al. 1995), albeit not over the same 1994–1998 time period. Superluminal speeds have also been found in optical observations of several other regions of the M87 jet, out to about 10″ (780 pc) from the nucleus,

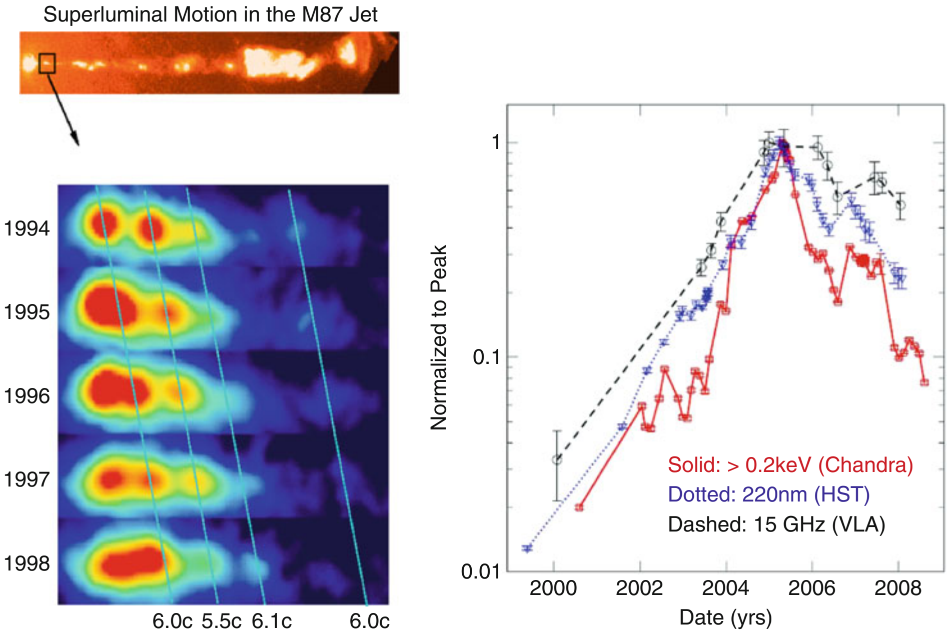


Fig. 7-22

Knot HST-1 in the M87 jet, located at a projected distance of 60 pc from the nucleus (see inset at *top left*), is the site of violent variability as well as superluminally moving components. The images at *left*, taken from Biretta et al. (1999, a color version of their Fig. 2), show the HST-1 region at five epochs, roughly once per year between 1994 and 1998. Four components are seen to move at speeds ~6c. As shown at *right* (taken from Harris et al. 2009, their Fig. 4), this region is also the site of violent variability; beginning in 2000, the knot began to increase in brightness, culminating in April 2005 when HST-1 had surpassed the core in flux by a factor of 2–6 in the UV and X-rays

with speeds that appear to decrease with distance from the nucleus, suggesting a gradual deceleration process may be at work. As of now no other radio galaxies have had such rapid motions detected, however a few other radio galaxies have been seen with mildly superluminal motions (1–2c), in particular B2 1144+35B (Giovannini et al. 2007, 1999).

Knot HST-1 is also the site of violent variability, another key property of blazars. As can be seen in ➋ *Fig. 7-22* (right-hand panels), beginning in 2000–2001, HST-1 began a massive increase in its flux, which culminated in March–May 2005 when for 10 weeks the knot was two to six times brighter than the nucleus (depending on the band) in the ultraviolet and X-rays. At maximum, the flux of HST-1 had increased by about 100 over its quiescent level. It is not known whether other regions in the large-scale jet of M87 also vary in brightness; however, the nucleus of M87 is also known to vary rapidly, both in the optical (Tsvetanov et al. 1998; Perlman et al. 2011) as well as in the X-rays (Harris et al. 2009).

Significant core variability has been seen in other radio galaxies, particularly Centaurus A, which is known to vary massively in the hard X-ray and millimeter bands (Jourdain et al. 1993; Beckmann et al. 2007; Abraham et al. 2007; Chitnis et al. 2009), as well as 3C 390.3 and 3C 120, both of which are known to be variable in the optical and X-rays. In those last two objects, there is, however, some controversy over the nature of the observed variability, as some of the properties are consistent with the less violent variability shown by Seyferts (both are broad-line objects; see Gliozzi et al. 2009 for 3C 390.3 and Marshall et al. 2009 for 3C 120), suggesting an origin outside the inner jet regions. However, both objects are known to exhibit apparent superluminal motion (see Gomez et al. 1998 for 3C 120, and Alef et al. 1988, 1996 for 3C 390.3), and in addition, Arshakian et al. (2010) and Abraham et al. (2007) argue strongly for the origin of the variability in 3C 390.3 in its jet. Radio variability is also seen for these objects as well as others in the radio (Ekers et al. 1983; Valtaoja et al. 1992), albeit in some cases mild enough that beaming may not be required for an explanation.

A well-known property of blazars is their high-energy gamma-ray emission – until recently, they were the only extragalactic objects known to emit at GeV and TeV energies, where in many cases it turns out that blazars emit the majority of their power (➲ *Fig. 7-23*). Blazar gamma-ray emissions are highly variable, and it turns out that these variations put strong, model-independent constraints on the compactness required, requiring relativistic beaming in their own right. Specifically, in order for gamma-rays to escape the source, the optical depth to pair production, $\tau_{\gamma\gamma}$, must be of order unity or less, which is equivalent to saying the compactness, a convenient dimensionless parameter that represents source luminosity divided by dimension, must be less than about 40 at the threshold for pair-production. That is, $\tau_{\gamma\gamma} = l/40 = 1$, where $l = (L/r)\left(\sigma_T/m_e c^3\right)$ is the compactness, with L and r being the source luminosity and dimension. The Thomson cross section, σ_T, is appropriate because most pairs will be produced by interactions with gamma-rays rather than particles (see ➲ Sect. 7). For 3C279 and PKS 0528+134, the first two blazars where this variable emission was detected, the inferred values for the compactness are 5,000–15,000, well in excess of the optical depth limit. To allow gamma-rays to be observed from these blazars, the true gamma-ray luminosity must be much smaller than observed and the true size much larger. Relativistic beaming has the effect that $L_{\text{obs}} \sim \delta^{(3+\alpha)}L$, where δ is the Doppler beaming factor (➲ 7.7) and α is the spectral index. If r is estimated from the variability timescale, then

$$l = \delta^{-5}\frac{L_{\text{obs}}}{\Delta t_{\text{obs}}}\frac{\sigma_T}{m_e c^4} \tag{7.7}$$

The limit $l \leq 40$ then translates to $\delta \geq 6$ for 3C 279 and $\delta \geq 7$ for PKS 0528+134, where L_{obs} has been evaluated at X-ray energies under reasonable spectral assumptions (Maraschi et al. 1992); similar limits are obtained for other gamma-ray blazars (Dondi and Ghisellini 1995). More recent observations (Abdo et al. 2009a) show much shorter timescale variability now in GeV observations, as small as days, and in TeV observations on timescales of less than an hour (Benbow et al. 2008). As shown in ➲ *Fig. 7-23*, M87 (Abdo et al. 2009b; Aharonian et al. 2006; Beilicke et al. 2005) is now known to emit in both GeV and TeV gamma-rays, with a luminosity much less than that seen in blazars, and is in addition, strongly variable on timescales as small as a day, while Centaurus A and NGC 1275 (Abdo et al. 2009c) are now confirmed as GeV emitters, with luminosity similarly orders of magnitude below that seen in blazars. Thus the detection of very-high-energy gamma-ray emission also argues for relativistic jets in both blazars and radio galaxies, and hence unification.

A final piece of evidence in favor of the unification of radio-loud classes of AGN can be found in their luminosity functions. Specifically, Urry and Padovani (1995) found that they could use a fit to the luminosity function of the 2 Jy radio galaxy sample and apply a simple beaming correction, and derive a reasonable fit to the luminosity function of both quasar-type and BL Lac-type blazars. That work requires Doppler factors for both populations of

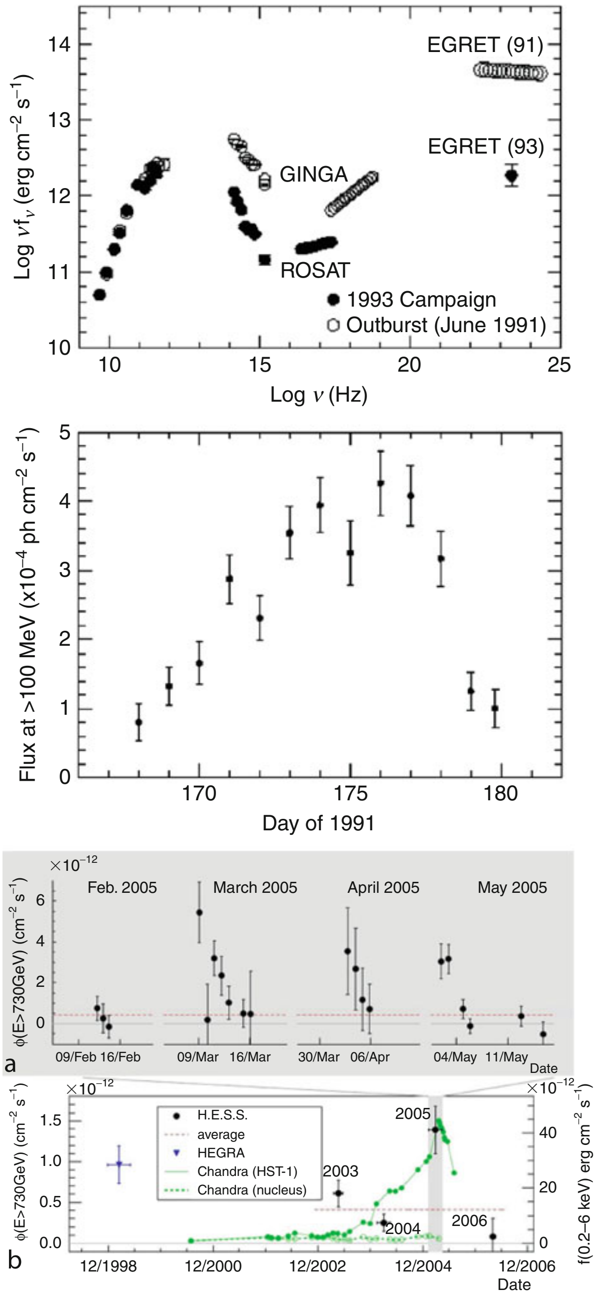


Fig. 7-23

At *top*, the spectral energy distribution for the blazar 3C 279 during 1991 and 1993, and at *middle*, the GeV variability seen by EGRET during the 1991 flare. At *bottom*, the fast variability now seen for M87 in the TeV band. Figures from Maraschi et al. (1994a), Kniffen et al. (1993), and Aharonian et al. (2006). The similarity between these variations is striking and in both cases requires relativistic beaming

between 2 and 40, with a power-law distribution, and critical viewing angles of 10–20°. More recent work (Padovani et al. 2007) has verified and strengthened this connection.

4 Nuclear Black Holes and Accretion

The development of the dominant paradigm for AGN occurred gradually. Of key importance was the realization that a black hole was necessary to drive the enormous luminosities and high-energy phenomena seen in AGN. The generation of large luminosities (as high as 10^{12-14} solar luminosities in quasars) in a region only a few light-days across is quite difficult. A stellar cluster, for example, would have to include many billions of stars in this small volume in order to explain these luminosities, and it is a simple matter to calculate that such a cluster would be unstable to collapse within very short timescales. It is for this reason that Donald Lynden-Bell (1969) first suggested the connection between active galaxies and black holes.

Accretion onto a black hole can be highly efficient in producing energy. If one assumes a small mass δm spiraling into a black hole from infinity, it will have an available (gravitational potential) energy of $\delta E = \frac{\mathrm{GM_{BH}}\delta m}{r}$, such that the upper limit on the luminosity is simply $L_{\max} = \frac{\mathrm{GM_{BH}}\dot{M}}{r}$. Note that this is an upper limit – not all of the gravitational potential energy can be extracted from infalling matter. If, however, one assumes that the matter goes in to 3 R_S (the location of the innermost stable orbit for a Schwarzschild black hole), then the efficiency ε is

$$\varepsilon \equiv \frac{L}{\dot{M}c^2} = \frac{\mathrm{GM_{BH}}\dot{M}}{(6\mathrm{GM_{BH}}/c^2)} \times \frac{1}{\dot{M}c^2} \approx 0.17 \tag{7.8}$$

A considerably higher efficiency can be reached if the matter goes all the way to the event horizon before plunging (as it would in a maximally rotating Kerr black hole). However, a counterveiling factor is that this calculation is purely Newtonian and does not take into account relativistic effects, nor does it take into account exactly how the energy is extracted or any fluid-dynamical effects within the accretion disk (about which more later). Taking these into account, the best estimate for ε is ~0.06 for the Schwarzschild case and 0.42 for the Kerr case. So, for example, if $\varepsilon \sim 0.1$, then a luminosity of 10^{46} erg s^{-1} requires an accretion rate $\dot{M} = 2M_\odot$ year^{-1}.

Another constraint on the energy that can be extracted from accretion onto a black hole comes from radiation pressure. A basic result from electromagnetism is that radiation exerts a momentum flux of $P_{\mathrm{rad}} = L/(4\pi r^2 c)$, and so there will be a force $F_{\mathrm{rad}} = \frac{L\sigma_T}{4\pi r^2 c}$ on each electron at a distance r, where σ_T is the Thomson cross section. So, if the radiative force equals that from gravity, the luminosity is

$$L_{\mathrm{Edd}} = \frac{4\pi G c m_p}{\sigma_T} M_{\mathrm{BH}} = 1.51 \times 10^{38} \left(\frac{M}{M_\odot}\right) \mathrm{erg\,s^{-1}}. \tag{7.9}$$

This result, known as the *Eddington limit*, is a result that is also important to X-ray binary systems as well as high mass stars. If the luminosity exceeds this quantity, radiation force must overpower gravity. It represents a fundamental limit to the luminosity of any accretion-powered source, as without matter streaming in, the accretion process cannot function (although, *n.b.*, there are a few objects where asymmetric, super-Eddington flows are suspected (Collin et al. 2006; Collin and Kawaguchi 2004)). This argument can also be inverted – that is, the large luminosities observed from AGN absolutely demand central black holes in the range 10^5–10^9 $M_\odot$, as if the central object were either less massive or not a black hole, the Eddington limit would be violated. In fact, if a black hole's accretion flow produces light with a fixed efficiency ε, its mass also sets a characteristic scale for the mass accretion rate – the Eddington rate:

$$\dot{M}_{\mathrm{Edd}} = \frac{L_E}{c^2 \varepsilon} = 3 \left(\frac{M_{\mathrm{BH}}}{10^8 M_\odot} \right) \left(\frac{\varepsilon}{0.1} \right)^{-1} M_\odot \,\mathrm{year}^{-1}, \tag{7.10}$$

An important – if obvious – point needs to be made here and that is that in order to reach high efficiencies ε, the matter cannot just quickly accrete into the black hole. If matter plunges into the black hole quickly and as discrete lumps (e.g., stars or planets), there is little prospect for extracting significant amounts of radiative energy from it. Thus, two critical parts of the equation are the time taken in the accretion process and whether accreted objects are disrupted into smaller bits as they stream in. The latter question is governed by tidal forces, and for a rough indicator of where these forces operate, one can turn to the Roche criterion, which expresses the competition between the smoothed out mass density of the central black hole and the gravitational forces (i.e., internal mean density) of the star or planet. For a self-gravitating object of mass M_* and radius R_*, this will mean that the object will be pulled apart by tidal forces when it approaches closer than the tidal radius,

$$r_t = R_* \left(\frac{M_{\mathrm{BH}}}{M_*} \right)^{1/3} \approx \left[\left(\frac{M_{\mathrm{BH}}}{10^8 M_\odot} \right)^{-2/3} \left(\frac{\rho_*}{\rho_\odot} \right)^{-1/3} \right] R_S, \tag{7.11}$$

where the latter scaling holds for an object at the solar density. For a solar-type star, this radius is *smaller* than the Schwarzschild radius when the black hole's mass is less than $10^8\ M_\odot$ – so that it is entirely conceivable that a star could accrete onto the black hole without being disrupted first. Thus, even though this radius varies weakly with the density of the object being accreted, it is important to realize that gravity alone will not serve to extract significant amounts of energy from matter. Therefore it is more profitable to consider the supply of matter at lower densities, that is from the interstellar medium. Furthermore, one also needs to consider the roles played by other processes – such as viscosity and disk structure – in the energy production mechanism. The actual efficiency of the process and hence the luminosity of any accreting black hole system are highly dependent on the type of accretion flow – disk, spherical, or otherwise – as well as whether the cooling within the disk is dominated by radiative or dissipative losses.

The data are decisively in favor of the association of AGN with nuclear black holes. A few AGN were included in the HST Nuker survey of nearby galaxies (Magorrian et al. 1998) for central black holes, in particular M87, where the data require a $3 \times 10^9 M_\odot$ black hole. This link has been strengthened by gas dynamical measurements of black hole masses for a number of other AGN (Wang et al. 2009; Onken and Kollmeier 2008; Zhang et al. 2008; Vestergaard et al. 2008; McGill et al. 2008; Hicks and Malkan 2008; Onken et al. 2007; Greene and Ho 2007; Kelly and Bechtold 2007). Moreover, reverberation mapping (➲ Sect. 5.1) has now been used to estimate the black hole masses for dozens of AGN (Vestergaard and Peterson 2006; Vestergaard and Osmer 2009; Bentz et al. 2009c) and work back to the M-σ relation (Ferrarese and Merritt 2000), which is found to be the same for AGN as it is for normal galaxies. Finally, masing molecular line features have been discovered in several Seyfert 2s. The data indicate that the material comes from warped, rotating disks a few thousand R_S from the black hole (e.g., ➲ *Fig. 7-24*). The advantage of this last method is that it is sensitive to only the accretion structure itself, as the resolution of the VLBI technique is a 100 times finer than that available from HST optical data.

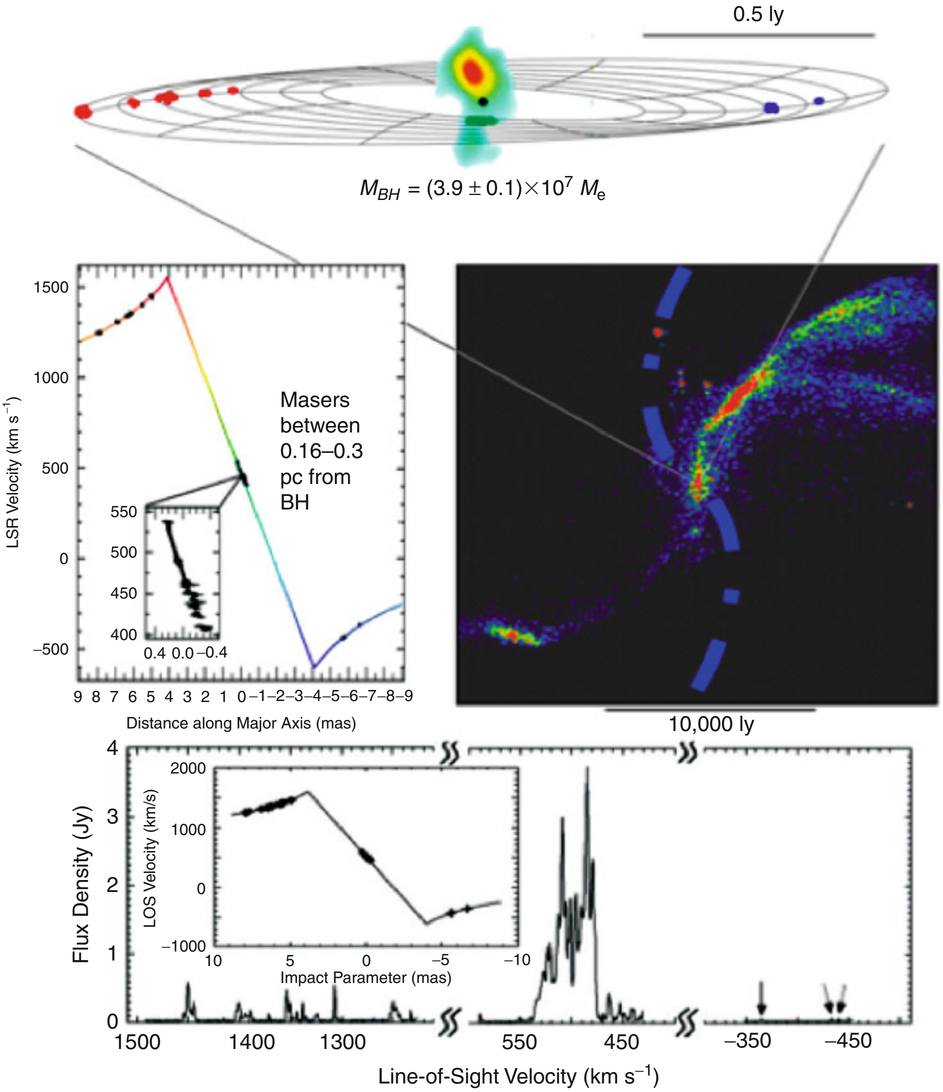


Fig. 7-24

Masing molecular features discovered in the outer accretion structure of NGC 1068. Included is a model fit for a Keplerian disk with a central *black hole* of the indicated mass (Figure taken from Greenhill et al. (1996))

None of these datasets, however, can tell us what kind of a black hole is present. The jury is still out on this; however, the latest data indicate that at least some AGN require near maximally rotating Kerr black holes. The reason for this is twofold. First of all, broadened Fe Kα lines have been observed in several AGN (e.g., *Fig. 7-25*), which not only require both special and general relativistic corrections to explain their profiles, but also require that material be present within 3 R_S, where the innermost stable orbit would lie for a nonrotating Schwarzschild

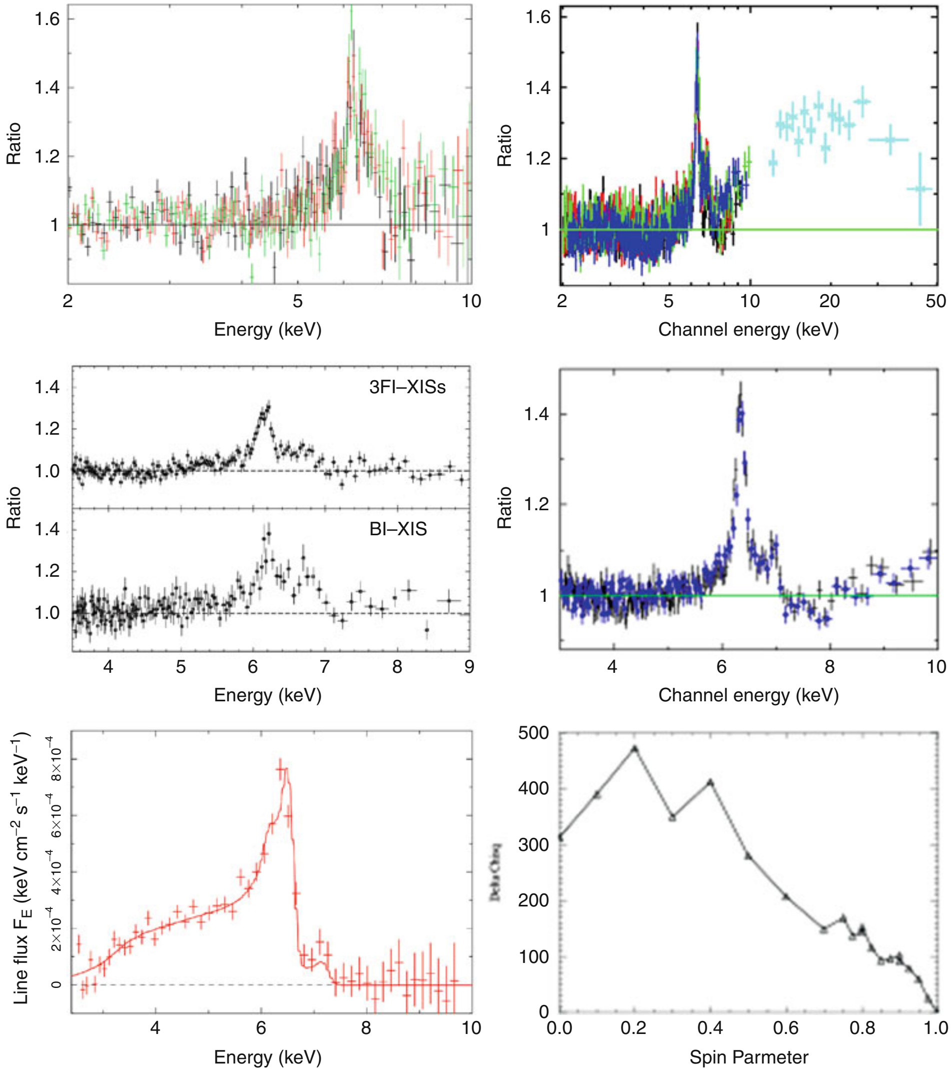


Fig. 7-25

X-ray spectra of four active galaxies showing the Fe Kα line profiles seen for objects best fitted by disks that truncate at 40 R_G (Mrk 335, *top left*, Larsson et al. 2007), 20 R_G (MCG -5-23-16, *top right* and *middle right*, Reeves et al. 2007), 10 R_G (3C 120, *middle left*, Kataoka et al. 2007), and <3 R_G (MCG -6-30-15, *bottom panels*, Fabian et al. 2002). This type of line profile can only be produced by the warped spacetime that exists close to the event horizon of a *central black hole*

black hole. The presence of cold material at smaller radii requires a rotating black hole, where the innermost stable orbit can move closer in to the ergosphere. If the black hole were not rotating, all of the material within 3 R_S would be in the "plunging region" where no stable orbit is possible, and in that case it would plunge into the black hole in a very short time with no further release of energy.

The second reason why rotating black holes may be required for many AGN comes from models of jet ejection (e.g., Meier et al. 2001; Tchekhovskoy et al. 2011; see ❯ Sect. 7), which indicate that the acceleration of a highly relativistic jet is much easier in the ergosphere surrounding a maximally spinning black hole. This latter work is not yet backed up by data, although some authors (notably Sikora et al. 2007) have linked the radio-loud/radio-quiet dichotomy to the spin of the central black hole.

4.1 Bondi Accretion

The simplest type of flow one can imagine into the central black hole is purely spherical. This is the classic problem of *Bondi accretion*. Imagine a black hole moving through a uniform medium of density ρ_∞ at some (relatively slow) velocity v. In spherical symmetry, the accretion rate from a radius r will then be simply $\dot{M} = 4\pi r^2 \rho_\infty v$. If the gas is adiabatic, then from a fluid-mechanical analysis of the surrounding medium, one can derive (see, e.g., Krolik 1999) that the overall accretion rate will be

$$\dot{M} = 4\pi\lambda\,(\gamma)\,\frac{(\mathrm{GM})^2}{c_s^3}\rho_\infty, \tag{7.12}$$

where c_s is the speed of sound and λ is a dimensionless parameter dependent only on the adiabatic exponent γ:

$$\lambda\,(\gamma) = \left(\frac{1}{2}\right)^{(\gamma+1)/[2(\gamma+1)]} \left(\frac{5-3\gamma}{4}\right)^{-(5-3\gamma)/[2(\gamma-1)]}. \tag{7.13}$$

This analysis then allows one to estimate the gas density necessary to feed an AGN of a given luminosity and mass:

$$\rho_\infty = 5\times10^{27}\left(\frac{L}{10^{45}\mathrm{erg\,s^{-1}}}\right)\left(\frac{T}{10^4\mathrm{K}}\right)^{3/2}\left(\frac{\varepsilon}{0.1}\right)^{-1}\left(\frac{M_{\mathrm{BH}}}{10^8 M_\odot}\right)\mathrm{g\,cm^{-3}}. \tag{7.14}$$

Bondi accretion represents the simplest solution to the problem of accretion in the sense that no assumption is made regarding the density of the material, incoming angular momentum, or the like. Of course more complex solutions, particularly disks (discussed next), are possible if the matter comes in with significant angular momentum.

As matter that is accreting spherically does not spend a large amount of time falling into the black hole, the majority of the regions around the black hole are optically thin, so that any emission is Bremsstrahlung, which has an emissivity proportional to $T^{-1/2}n^2$. An adiabatic accretion flow will then have a luminosity

$$L = \int_{R_s}^{r_{\mathrm{acc}}} 4\pi r^{1/2}\frac{AT_\infty^{1/2}}{\rho_\infty m_p^2}\left(\frac{\dot{M}}{4\pi}\right)^{7/3}\left(\frac{1}{2\mathrm{GM}}\right)^{7/6} dr. \tag{7.15}$$

$$= \frac{8\pi A}{m_p}\left(\frac{T_\infty^{3/2}}{\rho_\infty}\right)^{1/3}\left(\frac{\dot{M}}{4\pi}\right)^{7/3}\left(\frac{1}{2\mathrm{GM}}\right)^{7/6}\left(\frac{1}{R_S^{1/2}}\right). \tag{7.16}$$

After some manipulation, this simplifies to

$$L = \frac{8Ac}{4^{8/3}\pi m_p^{3/2}\mathrm{G}\sqrt{2k}}\left(\frac{\dot{M}^2}{M_{\mathrm{BH}}}\right). \tag{7.17}$$

For typical temperatures and densities, this emission is seen in the X-rays. The efficiency is then

$$\varepsilon = \frac{8A}{4^{8/3}\pi m_p^{3/2} G c\sqrt{2k}}\left(\frac{\dot{M}}{M_{\mathrm{BH}}}\right). \tag{7.18}$$

By putting in values typical of galactic nuclei, one sees that this is generally a very inefficient process. For example, Bondi accretion for Sgr $A*$ gives an efficiency $\varepsilon = 10^{-7}$, whereas most AGN require much higher efficiencies. As a result, Bondi accretion is probably important only for low-luminosity AGN, particularly LINERs and possibly also FR I radio galaxies (Allen et al. 2006). For more luminous objects, a different solution must be sought.

Thus spherical accretion, while likely common in galactic nuclei, actually has limited utility for fueling activity. In general, it is very inefficient in translating gravitational potential energy into radiative energy due primarily to the fact that the density of incoming material tends to be low and a given parcel will accrete onto the black hole relatively quickly (as opposed to disk accretion, for which see the next section).

4.2 Disk Accretion

One of the problems with spherical accretion is that the matter can fall into the black hole in a relatively short time, thus allowing little opportunity for the extraction of energy into radiative form. Also, it takes no account of the angular momentum of the matter streaming in. This latter consideration is important. If, for example, the central regions of the host galaxy are rotating, or if matter is streaming in to the central regions from a galactic interaction, then it will enter the central regions with considerable angular momentum, which must be dealt with if matter is to accrete onto the central black hole. In a typical galaxy, matter orbiting the nucleus at distances of kiloparsecs has a typical specific angular momentum of 10^{28}–10^{29} cm^2 s^{-1}, while matter orbiting at the last stable orbit of a $10^8 M_\odot$ black hole has a specific angular momentum four orders of magnitude less. Thus if rotation or large-scale streaming motions are important in a galaxy, it will be necessary for any matter that accretes to shed its angular momentum before being accreted onto the central black hole to fuel activity.

It is well known that typical galaxies are rotationally flattened into disk shapes by their angular momenta. It should therefore not be surprising that the same thing happens to accreting matter. Thus when angular momentum is important, one would expect material orbiting the black hole to stream into a disk, where as it loses energy and angular momentum (by heretofore unspecified processes) it will fall into a succession of ever-so-slightly smaller circles until finally it reaches the last stable orbit and plunging region. It should be mentioned that matter in this case will take a considerably longer period of time to make the journey from the nuclear interstellar medium into the black hole. This allows much more time for the extraction of energy, which was one of the problems plaguing Bondi accretion. Thus disk accretion can be much more efficient at extracting energy from accreting matter than Bondi accretion.

It is a good idea to begin our consideration of accretion disks with a discussion of their structure in terms of both temperature and density. If one assumes that the disk is supported against gravity by a pressure gradient in the vertical coordinate z, then from hydrostatic equilibrium,

$$\frac{dP}{dz} \cong -\rho\frac{\mathrm{GM}}{R^3}z \quad (z \ll R). \tag{7.19}$$

For an isothermal thin disk, the solution is an exponential

$$\rho = \rho\,(z = 0)\exp\left(-\frac{\Omega^2 z^2}{2c_s^2}\right) = \rho\,(z = 0)\exp\left(-z^2/h^2\right). \tag{7.20}$$

where the scale height is

$$h = \frac{\sqrt{2}cs}{\Omega} = \frac{\sqrt{2}csR}{v_\phi},$$

Ω is the angular velocity, and the velocity v_ϕ is just the Keplerian orbital speed for a given radius. This means that the specific angular momentum at any radius R will be $l = \sqrt{\mathrm{GM_{BH}}R}$. To flow inward, the gas must lose angular momentum through some unknown process. This can include redistribution of some kind (i.e., by gas at the smallest radii losing angular momentum) or a generalized wind or outflow. In either case, matter will be transported to small radii while angular momentum is carried out to a very large radius by some tiny fraction of the mass that must therefore leave the black hole's environs at large speeds.

If the gas is flowing inward, it will have a potential energy per unit mass of

$$E = -\frac{\mathrm{GM_{BH}}}{R}. \tag{7.21}$$

So for a small parcel of mass dM flowing inward by dR (see geometry in ➋ *Fig. 7-26*), its potential energy will change by

$$dE = \frac{\mathrm{GM_{BH}}}{R^2}dMdR. \tag{7.22}$$

Half of the energy will go into increasing the kinetic energy of the gas, while the other half is radiated. Therefore the luminosity at each radius will be

$$L = -\frac{\mathrm{GM_{BH}}\dot{M}}{2R^2}dR. \tag{7.23}$$

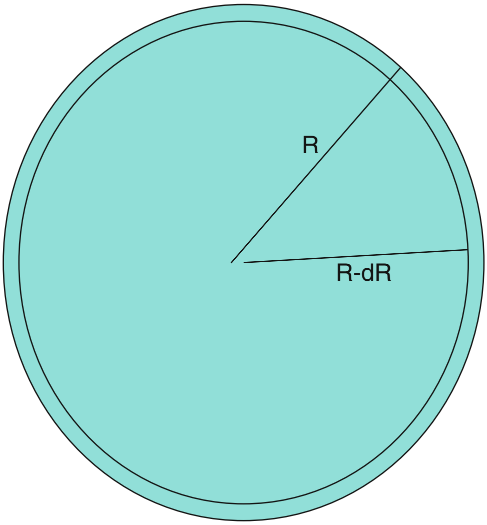


Fig. 7-26
Geometry for disk accretion

By dividing by the radiating area and equating the rate of energy loss assuming blackbody radiation, one can derive a radial temperature distribution for the disk:

$$T = \left(\frac{\mathrm{GM_{BH}}\dot{M}}{8\pi\sigma R^3}\right)^{1/4} = 6.8 \times 10^5 \varepsilon^{-1/4} \left(\frac{L}{L_E}\right)^{1/2} \left(\frac{L}{10^{46}\mathrm{erg\ s^{-1}}}\right)^{-1/4} \left(\frac{R}{R_G}\right)^{-3/4} \mathrm{K}. \tag{7.24}$$

It should be noted that this analysis, due to Shakura and Sunyaev (1973), does not account for the transport of angular momentum or boundary conditions (including general relativistic corrections near the black hole), and of course it assumes local thermodynamic equilibrium. A more correct analysis that includes these factors as well as the disk viscosity (see below) (Melia 2009) obtains

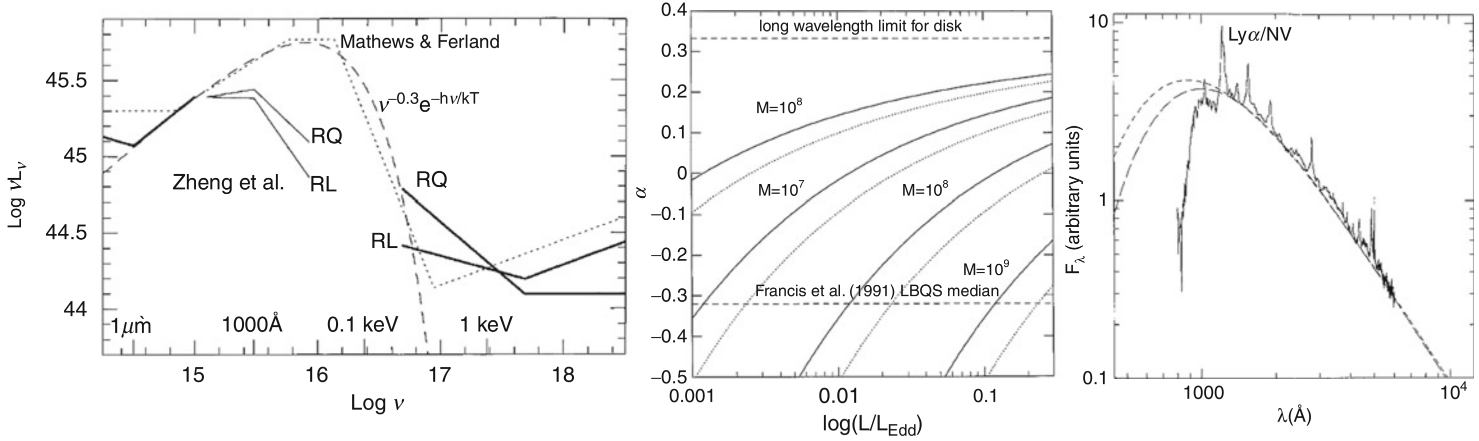
$$T = \left(\frac{3\mathrm{GM_{BH}}\dot{M}}{8\pi\sigma R^3}\left[1 - \sqrt{\frac{R_{\mathrm{in}}}{R}}\right]\right)^{1/4}. \tag{7.25}$$

The overall disk spectrum will then be an integral over the entire range of radii for which the disk exists, weighted by the mass in each annulus. The majority of the disk emission will come from the innermost region of the disk (i.e., the innermost few gravitational radii); however, as the accretion disk is expected to span at least two to three decades in radius, a similar range in temperature and hence frequency is to be expected from the disk spectrum. Theoretical models of disks indicate that the spectrum one should expect (e.g., Koratkar and Blaes 1999) is relatively flat, $\nu^{-1/3}$, over the range of temperatures seen in the disk, with an exponential cutoff at high frequencies. As shown by ➋ *Fig. 7-27*, this is consistent with the broad spectral energy distribution seen from most AGN in the near-IR, optical, and ultraviolet.

The exact mechanism for redistributing angular momentum is not clear. While it is clear this is a viscous process, it cannot simply be the microscopic viscosity, which is far too weak to have a noticeable effect. One way to produce a larger viscosity is via turbulence; however, there is no obvious reason why one might expect a turbulent disk as the Rayleigh stability criterion is satisfied for the Keplerian potential. It is true that since the interesting regions of AGN disks have temperatures ~10^5 K, He II/He III ionization transitions may occur at some radii, which would then be convectively unstable. However, such a mechanism is unlikely to exist throughout the disk, and, moreover, simulations appear to show that convection tends to carry angular momentum *in* rather than *out* (Stone and Balbus 1996). The conventional view now appears to be that magnetic fields are required to transport angular momentum outward, as required. In the presence of a magnetic field, the instability criterion is satisfied because the magnetic field couples fluid elements together. This mechanism, known as the magnetorotational instability (MRI) (e.g., Balbus 2003), can transport angular momentum outward. If the magnetic field grows as one goes to smaller radii, the MRI becomes stronger, and therefore one can see that this process is also intimately tied up with the mechanism for producing large-scale outflows and jets (about which more later), which also are likely responsible for carrying outward a fair fraction of the angular momentum of disk matter.

4.3 Disk Coronae and the 6.4 keV Iron Line

Accretion disks produce a quasi-thermal spectrum that tends to peak in the ultraviolet to soft X-rays. However, as already indicated, AGN usually exhibit emission that extends to much higher energies, ≥100 keV even for radio-quiet AGN. This emission cannot come from the disk



■ Fig. 7-27

Predictions and performance for accretion disk models. At *left* and *middle*, the optical/UV slopes predicted by simple models of disk accretion around Schwarzschild black holes, as a function of luminosity. The transfer function used is fully relativistic, and the *upper* and *lower* curves correspond to inclination angles of 66° and 26°, respectively. At *right*, a fit of a model accretion disk spectrum to the composite quasar spectrum of Francis (1996). Note the excellent fit, although the roll-off has yet to be replicated by data due to the difficulty of observations (All plots from Koratkar and Blaes (1999, their *Figs. 7-3*, *7-16* and *7-17*) itself or the broad- or narrow-line regions because the conditions are not right. However, one aspect not yet considered is whether, in a system where disk accretion is going on, the regions surrounding the accretion disk will contribute to the emissions observed. It is logical to assume that, at higher latitudes surrounding the accretion disk, there is an "atmosphere" or (as more often attributed) corona of hot gas. This corona can be much less dense than the accretion disk itself; however, thanks to well-known processes such as Compton scattering, it can account for a major portion of the AGN emission.

In Compton scattering, photons are scattered by electrons. It is important to point out that Compton scattering is a *mandatory* process – that is, if one has a region of space with a considerable population of free electrons, and a bright radiation source nearby, the process is unavoidable. These conditions are easy to satisfy in a lower-density region near the center of an AGN, and therefore it is critical to consider Comptonization and its implications for AGN – both in the corona as well as other regions (see ➋ Sect. 7 on jets, below). In this section, the theory of Comptonization and its impact on AGN are discussed. More information on the subject is available from, for example, Rybicki and Lightman (1986).

It is a well-known result from special relativity that when a photon is scattered by an electron, its wavelength changes by

$$\Delta\lambda = \frac{h}{m_e c}\,(1 - \cos\theta)\,, \tag{7.26}$$

and its final energy is related to its initial energy via

$$\Delta E = \frac{E^2}{m_e c^2}\,(1 - \cos\theta)\,. \tag{7.27}$$

In the electron's rest frame, the photon imparts energy to the recoiling electron. However, in the laboratory frame, it is possible for the electron to impart energy to the photon, up to $\Delta E = (\gamma - 1)\, m_e c^2$ (where γ is the electron's Lorentz factor) if the angle is right. This process, whereby the photon gains energy through scattering, is called inverse-Compton scattering. For low-energy electrons, this process is not important; it is only for high-energy particles (i.e., hot gas) where it takes on critical importance.

It can be shown that for a nonrelativistic, thermal distribution of electrons with temperature T_e, the average change in energy becomes

$$\langle\Delta E\rangle = (4kT_e - E)\,\frac{E}{m_e c^2}\,. \tag{7.28}$$

For a thermal distribution of electrons, a photon will gain energy if its energy $E \ll 4kT_e$, with the change in energy being proportional to γ^2. One factor of γ in this proportionality comes from the boosting of the photon into the electron's initial rest frame, while the other comes from the boosting of the scattered photon back into the lab frame. The process of inverse-Compton scattering can occur an arbitrary number of times (so long as the photon remains within the scattering region), as long as $E \ll 4kT_e$, with $\Delta E/E = 4kT_e\big/m_e c^2$. After N scattering events, the photon's final energy will be given by

$$E_f = E_i \exp\left(N\frac{4kT_e}{m_e c^2}\right) \tag{7.29}$$

Typically the number of scatters is parameterized in terms of the optical depth τ_{es}; in this parameterization, the average number of scatters will be $\max(\tau_{\mathrm{es}}, \tau_{\mathrm{es}}^2)$. Often one speaks in terms of the Compton y-parameter, where y is defined such that

$$y = \max\left(\tau_{\mathrm{es}}, \tau_{\mathrm{es}}^2\right)\left(4kT_e / m_e c^2\right) \tag{7.30}$$

and $E_f = E_i e^y$. If y is large – a case known as saturated Compton scattering – the average photon will reach the thermal energy of the electrons. This process is less important for accreting black hole systems. More important for AGN coronae is the unsaturated case, which is the suspected mechanism for producing the hard X-ray continuum in the broadband spectrum of AGN. In the context of accreting black hole systems, it is most convenient for the corona to speak of the *virial temperature* that refers to the average accretion energy per particle. Since the gravitational energy released per particle of mass m scales as $\mathrm{GM_{BH}}m/r$, which itself scales as $mc^2 / \left[R / \left(\mathrm{GM_{BH}}/c^2\right)\right]$, the virial temperature is independent of the size of the black hole. Electron virial temperatures of tens to hundreds of keV can be readily achieved in the innermost regions of black hole systems (the proton virial temperature is a factor ~2,000 higher), and this represents the maximum energy that Comptonized photons can achieve.

The resulting spectrum thus reflects the competition between the number of scatterings and the likelihood of these multiple scatters – although multiple scatterings become exponentially unlikely, they produce exponential energy gain. The two effects balance to some degree, and it can be shown that for a given Compton y-parameter, the resulting spectrum is a power law with photon index $\Gamma(N(y) \propto E^{-\Gamma})$ of approximately 1 (see Reynolds and Nowak 2003).

$$\Gamma = -\frac{1}{2} + \sqrt{\frac{9}{4} + \frac{4}{y}}. \tag{7.31}$$

As can be seen from *Fig. 7-13*, the typical hard X-ray spectrum of an AGN has a photon index of around 2, which means that the typical Compton y-parameter in black hole coronae is about 1. Of course the exact function $\Gamma(y)$ depends on the geometry and other assumptions, and a power-law form is only achieved for photon energies less than the electron thermal energy. As photons approach the thermal energy, they can no longer gain energy from scattering, so that the observed high-energy spectral cutoff – typically in the neighborhood of a few hundred keV – yields information about the temperature of the electrons in the AGN's corona.

It should be mentioned that the mechanism for heating the electrons to near-virial temperatures is currently unknown, although the leading hypothesis is that magnetic processes are dominant. Second, the geometry of the corona in AGN is completely unconstrained; various workers have produced models for a variety of geometries, as shown in *Fig. 7-28*.

A second set of processes, due to reprocessing in the accretion disk's outer layers of photons scattered within the corona and then "reflected" back toward the disk, are also important for AGN X-ray spectra. These processes may be modeled simplistically by supposing that the accretion disk is a semi-infinite slab of uniform density gas, irradiated from above by a continuum produced in the corona via thermal Comptonization. Hydrogen and helium are assumed to be fully ionized, but heavier elements are neutral. While a crude approximation, this is not far from the truth in a relatively "cold" accretion disk.

An X-ray photon coming from the corona can either be scattered by the free electrons associated with the ionized hydrogen and heliums, or the outer electrons of the other elements, or photoelectrically absorbed by one of the neutral atoms. For the latter process to occur, the photon's energy must be above the threshold energy for a given photoelectric transition. The transitions with the largest cross sections are those associated with the ejection of K-shell

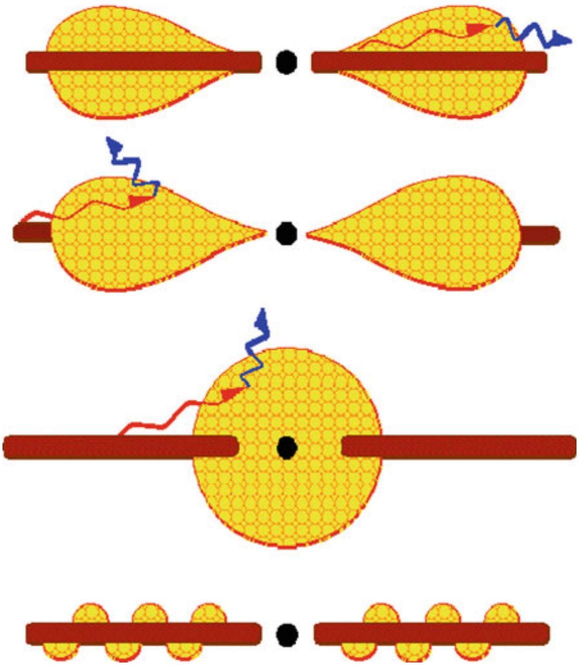


Fig. 7-28

Possible geometries for accretion disk-corona systems. The *top panel* is referred to as a "slab" or "sandwich," but it predicts spectra softer than observed. The remaining three geometries give a corona that is less effectively cooled by soft disk photons. The *middle two* are referred to as "sphere + disk" geometries, whereas the bottom geometry is called a patchy corona. (Figure taken from Reynolds (1996))

(i.e., $n = 1$) electron. Following K-shell photoionization, the resulting ion usually de-excites in one of two ways, both of which start with an L-shell ($n = 2$) electron dropping into the K-shell. This can occur with either the excess energy being radiated as a Kα line photon or a second L-shell electron can be ejected (often called autoionization or the Auger effect). Thus the output spectrum must include both of these effects: a "hump" due to the reflection at 30–40 keV and then a variety of Kα lines from different elements (see *Fig. 7-29*).

These features are indeed observed in most spectra, as shown by the recent Suzaku spectrum of MCG-6-30-15 (Reynolds and Miller 2009), as well as Chandra and XMM observations of several other AGN (Brenneman and Reynolds 2009). Importantly, one expects that the disk, and the associated corona, will truncate at a similar radius, namely, that of the innermost stable orbit (Reynolds and Miller 2009; Reynolds and Fabian 2008). Thus the shape of the observed Fe Kα line becomes a link to not only how far inward the disk goes but indeed what type of black hole the object hosts – as it is the spin parameter that determines the location of the innermost stable orbit (*Fig. 7-25*). However, in radio-loud objects, these features can be overshadowed by the continuum from the jet (which will be discussed in Sect. 7). In addition, there appears to be an inverse correlation between the 2–10-keV X-ray luminosity and the equivalent width of the Fe Kα line (Iwasawa and Taniguchi 1993; Nandra et al. 1997).

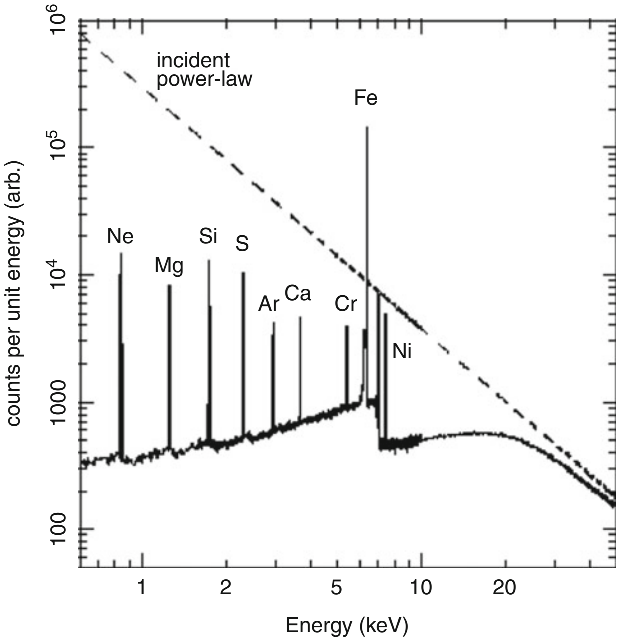


Fig. 7-29

Results of a Monte Carlo simulation demonstration of an incident power-law X-ray spectrum (shown by the *dashed line*) by a cold and semi-infinite slab with cosmic abundance (Figure from Reynolds (1996))

5 Emission Line Regions

As already discussed, one of the original characteristics by which AGN were first identified is their strong line emission in the optical. This is a property displayed by the great majority of all AGN, with the exception of BL Lac objects. The widths of these emission lines are generally interpreted as Doppler velocities, as the alternate interpretation of thermal broadening would require temperatures ~10^{10} K in the case of the broad lines. At that temperature, all atoms would be fully ionized so that no emission lines would be produced (in addition, a plasma at that temperature would efficiently produce e^+/e^- pairs, and the resulting annihilation line at 511 keV is not observed in the Gamma rays).

To get a first idea of the distance scale at which velocities ~10,000 km s^{-1} might be possible, one might suppose that the velocities are indicative of the local rotation around the central black hole. In this case, since

$$v_{\rm rot} \sim \sqrt{\frac{\rm GM}{R}} = \frac{c}{\sqrt{2}}\left(\frac{R}{R_S}\right)^{-1/2}, \tag{7.32}$$

one would expect a distance R ~ 500 R_S for velocities ~c/30. Hence, to a first approximation, the broad lines can be produced no closer to the black hole than ~1,000 R_S (although this estimate was based on the assumption of rotational motion, the infall velocity for free fall is within a factor 2). The region in which these broad emission lines are produced is called the *broad-line* *region* or BLR. By a similar argument, then, one would expect the narrow lines to be produced at greater distances from the central black hole, specifically at least 2 orders of magnitude more distance given that their widths are at least a factor 10 smaller.

5.1 Reverberation Mapping

A more direct way to measure the extent of emission regions in AGN and their geometry is provided by *reverberation mapping*. This technique, pioneered by B. M. Peterson (1993), utilizes the fact that heating and ionization of the BLR (for details see the ensuing discussion) are both accomplished by the central continuum source of the AGN, that is, the accretion disk. One would expect that variability in the accretion disk would therefore produce corresponding variations of the physical conditions in the BLR, and hence the broad lines observed. However, things are not quite as simple as that, as light (as with all other information) travels at a finite speed, namely c. If indeed a change in the continuum causes a change in the observed emission lines, then one would expect a delay $\Delta t \sim r/c$ between the change in the continuum and that observed in the emission lines – where r represents the distance between the regions where the two are produced. Thus what is done in reverberation mapping is that one correlates changes in a variety of emission lines to the changes in the UV continuum and thus infers the corresponding values of r, which correspond to the characteristic distances at which each of these lines are produced. One can also monitor different continuum bands to investigate the temperature structure of the accretion disk.

It should be mentioned that carrying out this technique is very demanding – one needs to continuously monitor the fluxes of both the continuum and lines (many of them) over a long period and then cross-correlate them against one another to look for correlated variability. The technique is illustrated in ➲ *Figs. 7-29* and ➲ *7-30*. Looking at the figure, it is important to point out the multiplicity of timescales that one needs to monitor: while the variations in the continuum and lines are on timescales of days to hours, the delays between the continuum and the broad lines occur on much longer timescales – months to a few years. Thus one needs to obtain data from many observatories in order to avoid gaps due either to bad weather or even the normal day-night cycle at single observing sites.

These campaigns (see ➲ *Fig. 7-31*) show that the BLR is typically at distances of light-months from the central black hole. Furthermore, the BLR extends over a wide range of radii, which consists of a variety of different layers; higher ionization potential lines are produced closer in to the black hole (and therefore have smaller delays observed) while lower potential lines are produced at greater distances from the black hole. For example, for the Seyfert 1 galaxy NGC 5548, for which ➲ *Fig. 7-30* shows the result of reverberation monitoring data, one obtains $r \sim 12$ light-days for $Ly\alpha$, about 26 light-days for C III], and about 50 light-days for Mg II. It should be noted that the finding that higher ionization potential lines are produced at smaller distances from the black hole is consistent with the observation that the higher ionization potential lines also tend to be broader than lines at lower ionization potential (Bentz et al. 2006, 2007). Interestingly, lines of higher ionization energy also tend to have a mean redshift that is systematically shifted blueward compared to narrower emission lines – thus hinting at a generalized outflow structure in which the BLR participates (see e.g., Risaliti and Elvis 2010). Another interesting finding from reverberation mapping data is that the extent of the BLR also scales with the luminosity of the AGN (Peterson et al. 2005; Kaspi et al. 2007; Bentz et al. 2009b) – larger BLRs are observed in more luminous objects and vice versa.

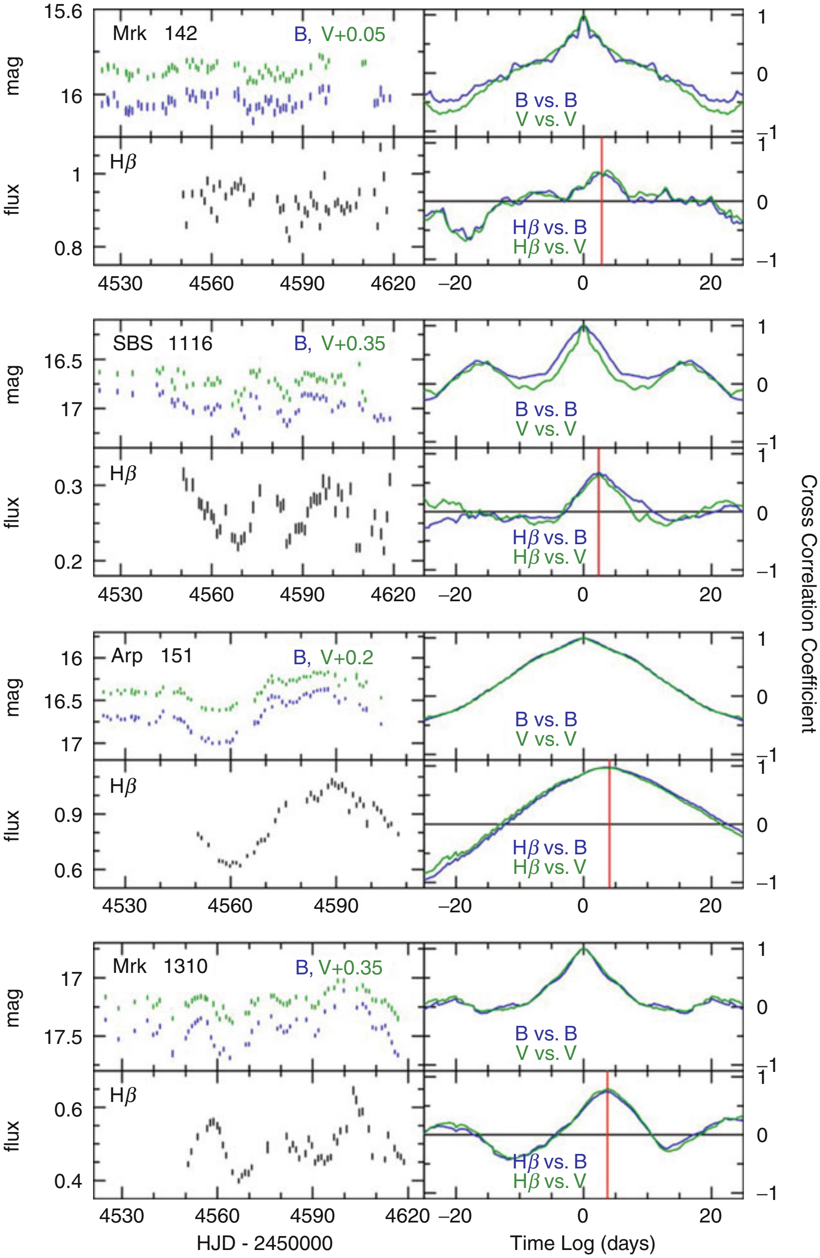


Fig. 7-30

***Left panels*: Photometric and Hβ *light curves* for Mrk 142, SBS 1116+583A, Arp 151, and Mrk 1310. The photometric measurements have units of Vega magnitudes, and the Hβ emission line fluxes have units of 10–13 erg s^{-1} cm^{-2}. *Right panels*: Cross-correlation functions for the *light curves*. For each object, the *top panel* shows the auto-correlation functions of the photometric *light curves* and the *bottom panel* shows the cross-correlation of Hβ with the photometric *light curves*. The *red vertical lines* mark the location of the measured lag time (Figure taken from Bentz et al. (2009a, their Fig. 1))**

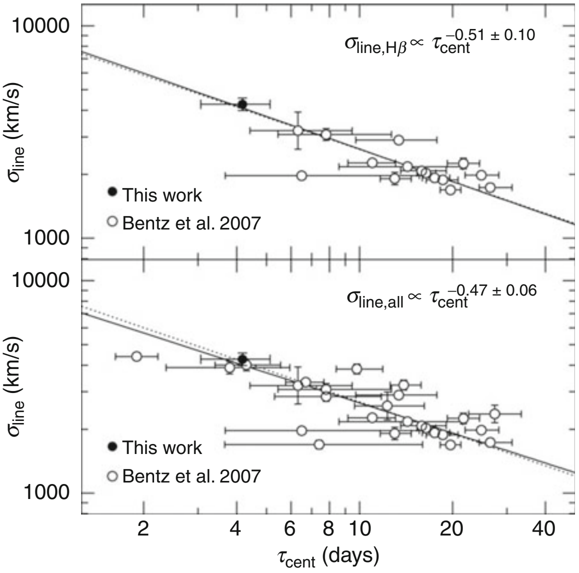


Fig. 7-31

Relationship between lag time and line width for several independent reverberation studies of NGC 5548. The *top panel* shows the relationship for Hβ reverberation results only, while the *bottom panel* shows the relationship for all broad emission lines with reverberation results. The *dark circle* in each panel is the Hβ result from this work, while the *open circles* are the compilation of results from Bentz et al. (2007) and references therein. The *solid lines* show the best fits to the relationship, with the slopes noted in each panel. The *dotted lines* show the relationship with the slope fixed at the value expected for a virial relationship, that is, −0.5

5.2 Physical Conditions in the BLR

The broad lines observed in AGN are nearly all permitted transitions typical of those seen in the warm interstellar medium. The BLR clouds are photoionized by energetic (i.e., UV and X-ray) continuum radiation from the AGN's accretion disk (See Sect. 5.4). This photoionization does not, however, represent the only heating source. There is clearly significant hydrodynamic heating as part of the generalized outflow that is taking place in the AGN, which the BLR clouds take part in. The main energy loss mechanism in the BLR clouds is emission of line radiation. Detailed photoionization models (e.g., Korista and Goad 2004, 2000; Korista et al. 1995; and references therein) are very successful at reproducing the line emission and line ratios seen in AGN (*Fig. 7-32*).

Those efforts use techniques similar to those used in ISM work (e.g., line ratios and the like) to fill in information regarding the temperature and density of BLR clouds as well as other information. What is found is that the typical densities in BLR clouds range from 10^7 up to 10^{11} cm^{-3}, near the critical density for CIII] λ 1909, thus making that line a very important diagnostic with higher values required for higher ionization potential lines, which (as detailed above) also appear to come from regions somewhat closer to the central black hole, while the lower values pertain to lower ionization lines. The typical temperature is around 20,000 K. Interestingly,

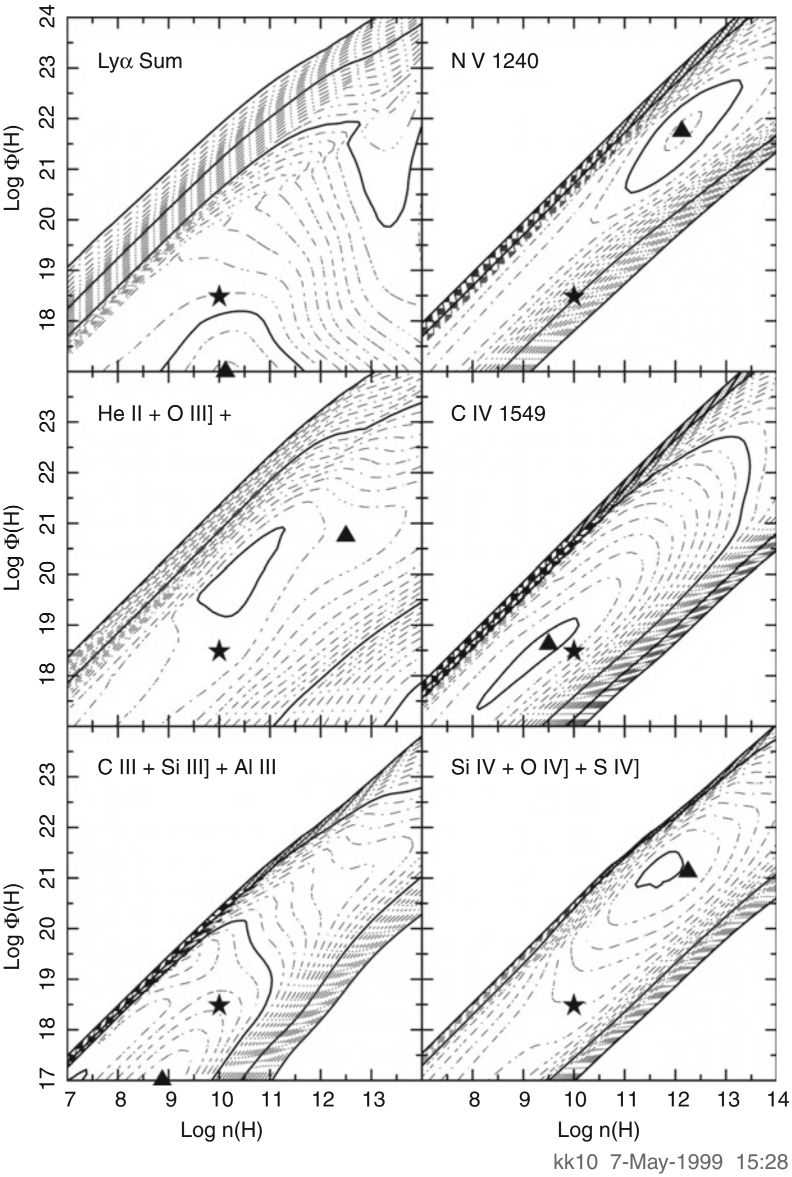


Fig. 7-32

Contours of log(EW) for six prominent UV emission lines or blends, referenced to the incident continuum at 1215A for full source coverage, are shown as a function of the hydrogen density and flux of hydrogen-ionizing photons. The total hydrogen column density within each cloud within the photoionization grid (generated using Cloudy; Ferland et al. 1998) is 10^{23} cm^{-2}. The EW is in direct proportion to the continuum reprocessing efficiency for that emission line. The smallest, generally outermost, decade contour corresponds to 1 Å; each solid line is 1 decade, and dotted lines represent 0.1 decade steps. The contours generally decrease monotonically from the peak to the 1 Å contour; the solid triangles mark the location of the peak of the dominant line within the blends (Ly α, He II, [C III], and Si IV). The solid stars are reference points marking the old "standard BLR" parameters. From Korista and Goad (2000)

the abundances seen in AGN broad-line regions do not appear to be typical of the interstellar medium. Arav et al. (2005, 2007) and Costantini et al. (2007), for example, find typically ~2× solar abundances of C, N, and O in the outflow and broad-line region of Mrk 279, using data from HST, FUSE, and Chandra. The origin of these supersolar abundances is unclear, but it hints at links between the host galaxy (i.e., the number of stellar generations in an AGN nucleus) and the presence of AGN.

There is considerable controversy about the filling factor and number of clouds required for BLR. While HST observations (e.g., Gabel et al. 2005; Scott et al. 2009, among others) suggest a filling factor ~0.1 and a complex kinematic structure within the BLR, high-resolution spectroscopy (Arav et al. 1997, 1998) show that an extraordinarily high number of clouds – upwards of 10^7 and possibly upwards of 10^8 – are required to explain the smoothness of the broad-line profiles observed in bright, albeit low-luminosity Seyfert AGN such as NGC 4151 and Mrk 335. These two results are at least outwardly difficult to reconcile with one another; however, it is possible for both to be true in a turbulent region. However, the combination of this large number of clouds and filling factor ~0.1 indicates that the discrete clouds that make up the BLR must be relatively small, $\sim 10^{14}$ cm in size at most. This small size has significant implications for the stability (i.e., evaporation time) of BLR clouds, especially under these conditions.

5.3 Physical Conditions in the Narrow-Line Region

In addition to the broad emission lines discussed above, one also observes a variety of narrow-line features. Their typical line widths are ~500 km s^{-1}, considerably narrower than the lines of the BLR. By analogy to the BLR, the region in which these lines are produced is known as the *narrow-line region* or NLR. The strongest line from the NLR is, besides $Ly\ \alpha$ and C IV, the forbidden [O III] line at 5007. The existence (and indeed dominance) of forbidden lines implies that the gas densities in the NLR are significantly lower than in the BLR.

Like the BLR, the NLR gas is assumed to be photoionized by the UV and X-ray continuum from the central engine. Photoionization modeling is described in detail in ➋ Sect. 5.4. From estimates analogous to those used for the BLR (see above), one can obtain the physical conditions of the NLR. An example of this work is shown in ➋ *Fig. 7-33*.

It should be noted, however, that one cannot apply the technique of reverberation mapping, as no correlated variability has been observed, and the extent of the NLR is believed to be of order 100 pc. The line ratios of allowed and forbidden lines yield typical electron densities $n_e \sim 10^3\text{–}10^5\ \mathrm{cm}^{-3}$ for the gas where the lines originate. The typical temperature is ~15,000 K, similar to but perhaps slightly lower than the BLR, and the filling factor for the line-emitting material is significantly less than one – typically in the neighborhood of 10^{-2}. Hence one can use a similar geometrical picture as for the BLR, namely, that of line-emitting clouds, perhaps embedded in a generalized outflow.

Since the NLR is much more extended than the BLR, in nearby objects one can hope to resolve it with modern telescopes. An example of such data is seen in ➋ *Fig. 7-34*. As can be seen, the morphology of the NLR is highly interesting – rather than spherically symmetric, it appears to show two cone-shaped regions on opposite sides of the nucleus. Thus it would appear as if the ionization of the NLR by the AGN's continuum radiation is not isotropic at these scales, but instead depends strongly on the direction. This is an observation which agrees strongly with unified schemes (➋ Sect. 6).

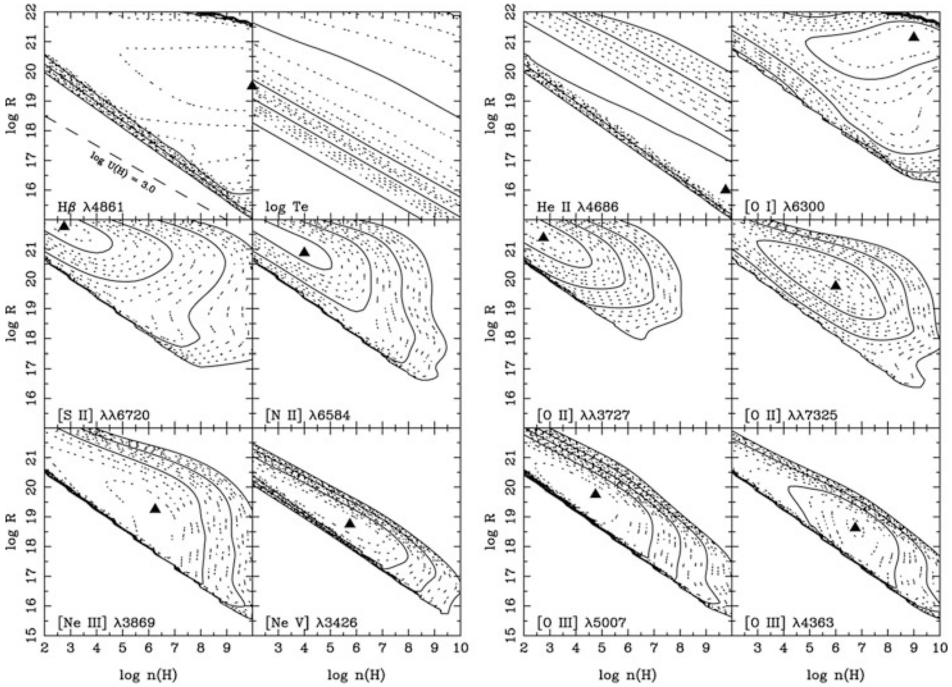


Fig. 7-33

Contours of constant logarithmic line EW as a function of log *R* (distance from the ionizing AGN, in cm) and log *n*(*H*) for the 23 emission lines indicated, referenced to the incident continuum at 4,860 Å. The cloud distance from the central ionizing source, *R*, was normalized assuming a log(L_{ion}) = 43.5, and the grid computed using Cloudy (Ferland et al. 1998). The triangle is the location of the peak in the equivalent width distribution, and the contours decrease downward to the outer value of 1 Å. All points within the grid assume full source coverage. The *upper right-hand* plot in panel a is the log(T_e) at the illuminated face of the cloud. The temperature decreases from 10^7 K in the *lower left-hand corner* of the plot to 10^3 K in the *upper right-hand corner*. The *bold lines* represent 1 dex increments, and the *dotted lines* are 0.2 dex steps, for all panels (Figure taken from Ferguson et al. (1997))

5.4 Photoionization

As discussed above, the temperatures in the emission line regions of AGN are typically of order 10,000–20,000 K. How, then, do the regions around the central engine achieve a large enough abundance of ionized species to generate emission lines of highly ionized species, including [C IV, O III], etc.? Collisional ionization, which operates in supernova remnants and some other nebulae, cannot provide the answer: under pure shock-wave heating, the [O III] lines would require temperatures in excess of 5 × 10^4 K, and even higher for some of the higher-ionization lines. In addition, one expects a relationship between the temperature and degree of ionization, which is not observed. What is seen instead is a relatively constant temperature in both narrow

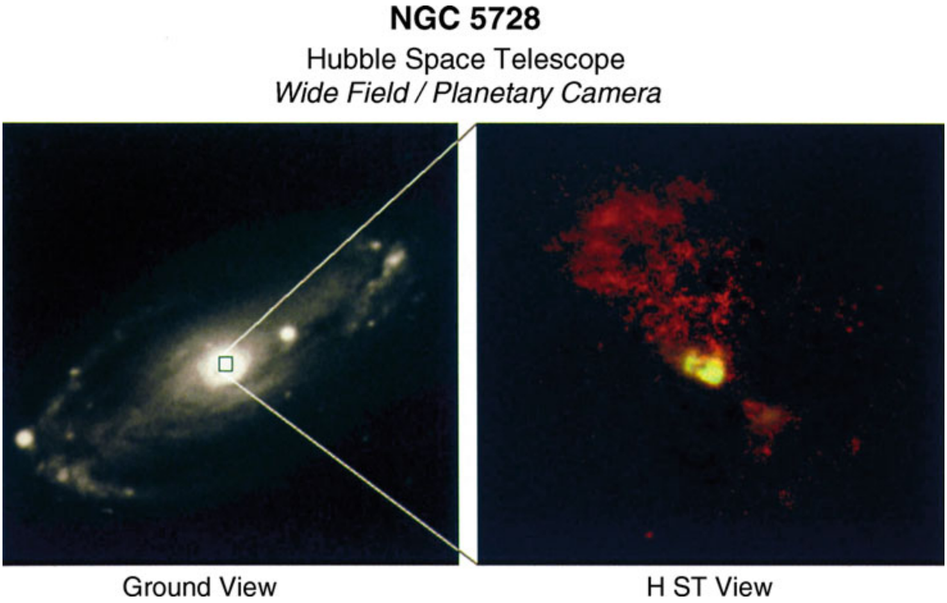


Fig. 7-34
Images of the Seyfert galaxy NGC 5728, showing its remarkable ionization cones. At *left*, a ground-based image of the galaxy. At *right*, a composite false-color image made from HST observations taken in the Hα+ [NII] and [OIII] lines (Wilson et al. 1993). As can be seen, the ionized gas lies in two conical regions located on either side of the active nucleus; these regions represent the narrow-line region gas in this object

and broad emission line clouds, with a wide range of densities. Reverberation mapping campaigns also show that the highest ionization broad emission line clouds lie closest to the active nucleus (➋ *Fig. 7-31*) – precisely where the flux of ionizing, high-energy radiation would be the highest. Moreover (➋ Sects. 2 and ➋ 4), the active nucleus is a copious source of high-energy UV and X-ray emission, with a roughly power-law shape. Thus these considerations lead to the preeminence of photoionization as the primary excitation mechanism in both the broad and narrow emission line regions.

The physical balance in the broad- and narrow-line regions is dictated by the following concerns. Energy input to the emission line regions can come from both photoionization as well as hydrodynamic processes. The latter can include both generalized outflows (such as those in the model of Risaliti and Elvis 2010), as well as any interaction with the jet (in radio-loud objects) or obscuration lines. The primary method for energy output is via radiative cooling by collisionally excited lines, which increases rapidly with temperatures above 10,000 K, which is the temperature where hydrogen becomes dominantly ionized across a wide range of physical conditions. This fact tends to keep the temperature in the 10–20,000 K range over a wide range of input ionizing continua.

It is usually assumed that the photoionized gas is far enough from the central source that the gas can be modeled as an infinite slab. A given model will generally assume a constant density in the slab, with the gas in local thermodynamic equilibrium. The Boltzmann and Saha equation are solved at each point, thus allowing the ionization and thermal structure to be determined within the gas along a radial direction from the central source. This will produce a depth-dependent opacity at every frequency as well as emission coefficients for different lines, allowing one to solve the equation of radiative transfer given the input emission spectrum. Several photoionization codes are in use; the most commonly used are CLOUDY (Ferland et al. 1998), Ion (Netzer 1990), and Xstar (Kallman and Bautista 2001).

The calculations these models do are quite complex, but the results can be encapsulated in two physical quantities: the number density and state of ionization. The number density is usually expressed in terms of the hydrogen column, whereas the ionization state is a more complex quantity, usually parameterized in terms of the *ionization parameter*, defined as the dimensionless ratio of the ionizing photon density to the electron density. In its simplest form – where ionization is done only by incident radiation – the ionization parameter can be expressed as

$$U = \frac{1}{4\pi r^2 c n} \int_{\nu_o}^{\infty} \frac{L_\nu}{h\nu} d\nu, \tag{7.33}$$

Here L_ν is the ionizing luminosity per unit frequency interval above the Lyman limit (i.e., $h\nu_0 = 13.6$ eV), and n is the number density of the gas in the slab. This formalism can be extended fairly trivially to the case of collisional ionization by specifying a "pressure ionization parameter," usually written as $\Xi = L/\left(4\pi r^2 c p\right)$, where p is the gas pressure.

The methodology one adopts is that one varies the density and ionization parameters until the predicted line luminosities match those observed. One must incorporate into the code all relevant atomic data, including recombination rates, ionization cross sections, charge exchange rates and the like. For each line, the luminosity is calculated, assuming that the total number of ionizing photons emitted by the central source must balance the number of recombinations in the ionized gas. These are of course related directly to the total number of line photons emitted in the gas. Thus, for any line, the equation of radiative transfer can be written (e.g., for Hβ)

$$L_{\mathrm{H}\beta} = h\nu_{\mathrm{H}\beta} \frac{\alpha^{\mathrm{eff}}_{\mathrm{H}\beta}\left(H^0, T\right)}{\alpha_B\left(H^0, T\right)} \frac{\Omega_{\mathrm{ELG}}}{4\pi} \int_{\nu_0}^{\infty} \frac{L_\nu}{h\nu} d\nu, \tag{7.34}$$

where Ω_{ELG} is the solid angle covered by the emission line gas (so that $\Omega_{\mathrm{ELG}}/4\pi$ is the covering factor), $\alpha^{\mathrm{eff}}_{\mathrm{H}\beta}\left(H^0, T\right)$ is the effective recombination coefficient for the Hβ line, and $\alpha_B\left(H^0, T\right)$ is the recombination coefficient for H^0, such that the ratio of these two coefficients is the number of Hβ photons produced per hydrogen recombination. See, for example, Osterbrock and Ferland (2006) as well as Crenshaw et al. (2003) for detailed descriptions of the process and assumptions of photoionization models.

Thus once one has measured the input photon spectrum as well as the luminosity in each of the emission lines, one can calculate parameters for the regions where each line is generated using (➲ 7.33) and (➲ 7.34), with the former modified as appropriate for the line in question. It is the output of these calculations that has been shown in ➲ *Figs. 7-32* and ➲ *7-33*. Note also that this same procedure is used to model the ionization structure of any intrinsic absorption region (➲ Sect. 6), where the models are iterated in density and ionization parameter to attempt to match the observed ionic column density.

Once these calculations have been done for each emission line, it is also useful to comment on the total mass of the BLR, which allows us to close the circuit begun at the beginning of this section by discussing the mean radius of these regions. The luminosity in a given emission line can be rewritten as

$$L\,(\mathrm{H}\beta) = n_e n_p \alpha^{\mathrm{eff}}_{\mathrm{H}\beta} h\nu_{\mathrm{H}\beta} V \frac{\Omega_{\mathrm{ELG}}}{4\pi}, \tag{7.35}$$

where V is the volume of the line emission region. Then, if one assumes roughly solar abundances, it can be shown that $n_e \approx 1.5 n_p$, and the mass of the region can then be calculated easily, as can its volume. The NLRs of the most luminous Seyfert galaxies have $L(\mathrm{H}\beta) = 2 \times 10^8 L_\odot$, which gives a total mass $M_{\mathrm{ion}} \approx 7 \times 10^5 (10^4/n_e) M_\odot$ if a roughly spherical NLR is assumed. Of course, this very elementary assumption does not pay attention to the underlying geometrical complexity of the emission line regions (e.g., ➲ *Fig. 7-34* for the NLR and ➲ *Fig. 7-31* for the BLR); however, it is adequate to an order of magnitude.

6 Nuclear Obscuration: Tori, Broad, and Narrow Absorption Lines

As shown by ➲ *Fig. 7-18*, the basic picture underlying the unified scheme places a geometrically and optically thick cloud of gas, often described as the torus surrounding the broad-line regions along the equatorial plane. This obscuring region is central to our view of AGN – when viewing the AGN along an equatorial line of sight, the torus obscures the broad-line emission, but the narrow emission line region is still visible, resulting in a type 2 object. In comparison, a pole-on view would provide a direct view of both the narrow and broad emission line regions, resulting in a type 1 classification. The presence of the torus can account for the ionization cones commonly observed in AGN in both emission lines and polarized flux through shadowing of the ionizing radiation by the torus, resulting in the biconical shape shown in ➲ *Fig. 7-34* (e.g., Packham et al. 1997). The torus also accounts for the X-ray differences, where soft X-rays are at least partially absorbed during their passage through the torus (Maiolino and Risaliti 2007). In addition to the torus, there are other absorption regions in AGN, from which one observes a variety of line features. Both will be discussed in this section.

6.1 The Torus

By far the most ubiquitous and well known of these obscuration regions is the torus, which as already discussed is central to unified schemes. Before torus models are discussed in detail, it is important to note that torus models currently rank as the least well constrained of all the different parts of the unified scheme simply because of how recently good data became available. Classical models of the torus assumed for simplicity a uniform dust distribution within a large, extended obscuration region (e.g., Pier and Krolik 1992; Granato and Danese 1994; Efstathiou and Rowan-Robinson 1995). However, these models require large (100 pc scale) tori to produce long-wavelength emission, and fine-tuning to account for differences between types 1 and 2 sources. However, the mid-IR observations described in ➲ Sect. 2.5 have thrown these notions largely into disrepute, as such a large, extended torus would easily have been resolved in nearby sources with ground-based observations. Moreover, such models fail to replicate the silicate feature characteristics and the extremely large ($N_H > 10^{24}$ cm^{-2}) column densities that X-ray data indicate for many sources.

More modern models of the torus must cope with size limits of only a few parsecs from the mid-infrared, as well as reverberation mapping observations (Suganuma et al. 2007) which place the 1–2 μm emission region just outside the broad-line region, that is, typically at distances of a few tenths of parsecs in Seyfert galaxies and more distant in more luminous objects, in agreements with simple calculations for the dust sublimation radius. These results are difficult to explain under homogeneous, extended torus models, but are consistent with the new breed of compact, but patchy torus models whereby the distribution of dust is clumpy (Elitzur et al. 2004; Elitzur 2007; Elitzur and Shlosman 2006; Nenkova et al. 2008a, b; see ➋ *Fig. 7-35*). The fundamental distinction of the inhomogeneous density distribution is that radiation can freely propagate between different optically thick clumps.

The model formalism accounts for both direct heating by the AGN and indirect heating by the ambient clouds' emission, so some dense clouds remain cool to provide long-wavelength emission within a compact volume. It is relatively elementary to derive that under these assumptions, the intensity at a point s generated by clouds along a given ray is

$$I_\lambda^C(s) = \int^{s} e^{-t_\lambda(s',s)} S_{C,\lambda}(s')\, n_C A_C(s')\, ds', \tag{7.36}$$

where $t_\lambda(s',s) = N(s',s)\left(1 - e^{-\tau_\lambda}\right)$' $N(s',s) = \int_{s'}^{s} n_C A_C(s)\, ds$ is the mean number of clouds between s' and s; n_C and A_C, respectively, are the number per unit volume and area of clouds along s; and τ_λ is the optical depth at wavelength λ, and all clouds are assumed to be identical. This is an exact analog to the general solution of standard radiative transfer in continuous media. If radiation is propagating from s′ to s, then it will have a probability of escaping that is equal to

$$P_{\text{esc}}(s',s) = e^{-t_\lambda(s',s)}. \tag{7.37}$$

Thus the only difference between clumpy and continuous media is that by integrating, one replaces the standard optical depth with its equivalent t_λ^s and the absorption coefficient is replaced by the product $n_C A_C$. Because of the particulate composition of matter, this equation is always valid in a statistical sense only, corresponding in principle to the intensity averaged along the same path over an ensemble of many sources with identical average properties.

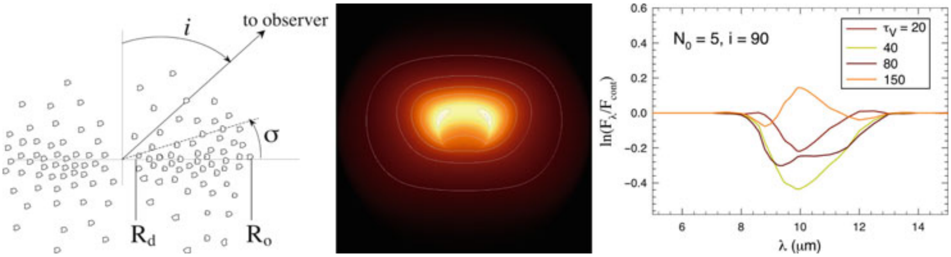


Fig. 7-35

Left*: In the clumpy torus model, the clouds follow a power-law distribution with radius from the inner radius, R_d, to outer radius R_o. The clouds are concentrated in the equatorial plane, distributed with scale height *s*, which has a Gaussian edge (Elitzur et al. 2004). *Center*: Simulated image of NGC 1068 at 8.8 μm, with logarithmically spaced contours. The torus extends horizontally, but on the smallest scales; the images extend vertically because of optical depth effects. *Right*: Continuum-divided simulated spectra show a range of behavior of the 10-μm silicate feature, even at fixed viewing angle, *i

To integrate a path containing a background source, such as the line of sight to the AGN, requires different handling since one cannot then average. For each line of sight, there are k intervening clouds, with Poisson probability P_k, and one then generates tabulations of intensities and their associated probabilities (I_k, P_k) with an actual source corresponding to one particular member of this probability distribution. Thus if the normalized spectral shape of the AGN radiation is $f_{e\lambda}$, the fraction of the AGN luminosity that emerges through a spherical shell of radius r centered on the nucleus is, on average,

$$p_{\mathrm{AGN}}(r) = \int_0^1 d\sin\beta \int d\lambda f_{e\lambda} P_{\lambda,\mathrm{esc}}(r,\beta). \qquad (7.38)$$

where $P_{\lambda,\mathrm{esc}}(r,\beta)$ is the probability for a photon of wavelength λ emitted by the AGN in direction β to reach radius r. Therefore, the fraction of type 2 sources in a given sample will be $f_2 = 1 - p_{\mathrm{AGN}}(R_{\mathrm{out}})$ where R_{out} is the outer torus radius.

Within the clumpy models, one generally assumes a power-law distribution of clouds with radius from the inner radius, R_d (presumably where dust sublimates), to outer radius R_{out}. One general result is that there is no such thing as a single-temperature torus – the distribution of temperatures in a given cloud will be dependent on the optical depth within it, with the coldest dust temperatures found at the highest optical depth. Thus in this context it is important to understand that the radius for the 2-μm emission found by Suganuma et al. (2007) does not represent the inner radius of the torus but rather the mean distance of the torus clouds themselves because what Suganuma et al. would have measured would be just the dust in the outermost layers of clouds located in the torus, not necessarily those located in a certain part of the torus.

➋ *Figure 7-36* shows the spectral energy distribution of slabs illuminated by AGN radiation in the case of two different observer directions and a number of temperatures. As can be seen, the general features of AGN mid-IR continua (➋ Sect. 2.5) are duplicated. Thus the clumpy models capture the broad spectral distinction of type 1 and type 2 AGN, with only the former affording direct views of the central engine and remaining bright across the entire UV-FIR bandpass, while producing nearly isotropic MIR emission. Most significantly, the models can reproduce the variety of observed MIR spectral characteristics, notably weak silicate absorption and emission, and a range of continuum slopes. The observations therefore provide powerful diagnostics of these physical conditions in the galaxies' central regions. Also, the emergent IR flux scales directly with the intrinsic luminosity of the AGN. Fitting a model to an observed spectrum therefore immediately reveals the power of the central engine, which is not otherwise evident in the obscured (type 2) cases.

6.2 Line Absorption

In addition to the above, AGN spectra contain a variety of line absorption features. This section will discuss only absorption that is *intrinsic* to either the AGN or its host galaxy rather than caused by intervening gas that lies by chance along the line of sight. While the latter features, seen in all AGN spectra, are interesting in their own right and place fundamental constraints on the physical state of the intergalactic medium and high-redshift galaxies, they are not germane to the subject of AGN physics.

Almost all varieties of AGN exhibit absorption, but its character seems to vary widely from one class to the next – both broad and narrow features are seen, with a large range in properties

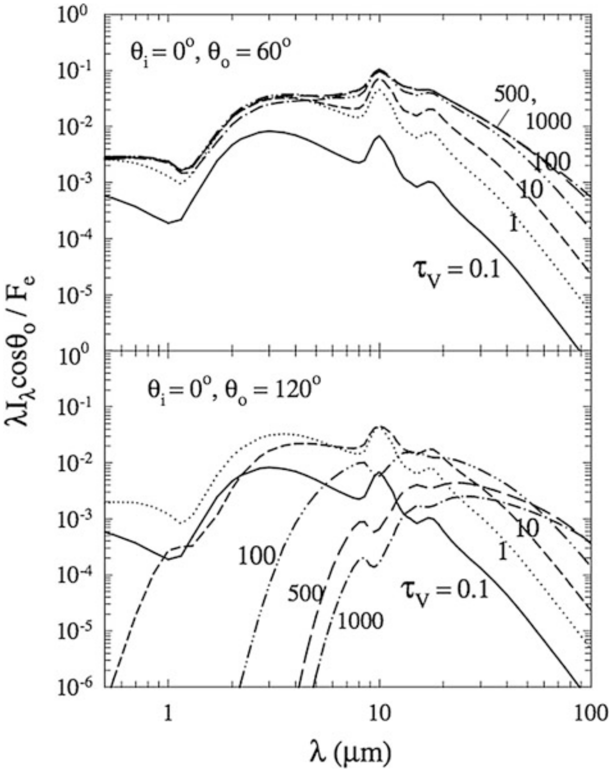


Fig. 7-36

Optical depth dependence of the SED of clumpy slabs illuminated by normal radiation to a maximum temperature of 850 K. In the *top* and *bottom* panels, the observer direction is 60° from slab normal on the illuminated and dark sides, respectively (Nenkova et al. 2008a, their Fig. 8)

that is at some level luminosity dependent. However, one feature appears to be common among all of these absorbers, namely, that any velocity shift seen is toward the blue (as compared to the AGN's systemic velocity) – that is, the absorbing material is approaching us and must therefore be moving *away* from the AGN. This last fact will be discussed in detail at a later time; however, this discussion makes clear that these absorption lines offer a view into material that is flowing out from the active nucleus, driven by some process that is connected with the overall energy generation mechanism and hence fundamental to our understanding of the AGN itself.

The features observed in AGN spectrum range from features that are narrow (tens to hundreds of km s^{-1} in width) to extremely broad (from 1 to 50,000 km s^{-1}). The former are found in a majority of Seyfert galaxies, both in the optical/UV (e.g., *Fig. 7-37*) and X-ray (*Fig. 7-14*), as well as in many quasars. When observed in Seyferts, these features have typical widths of 20–400 km s^{-1}, while their more luminous cousins tend to have broader features, up to ~1,000 km s^{-1}. The blueshifts of these features range from ~0 to ~2,000 km s^{-1} in Seyferts and up to 5,000 km s^{-1} in more luminous quasars. These lines can be detected with only moderate spectral resolution and signal to noise, given their high column depth. Typically, these features are seen in C, N, and O in electric dipole transitions from the ground to high states – such as in the

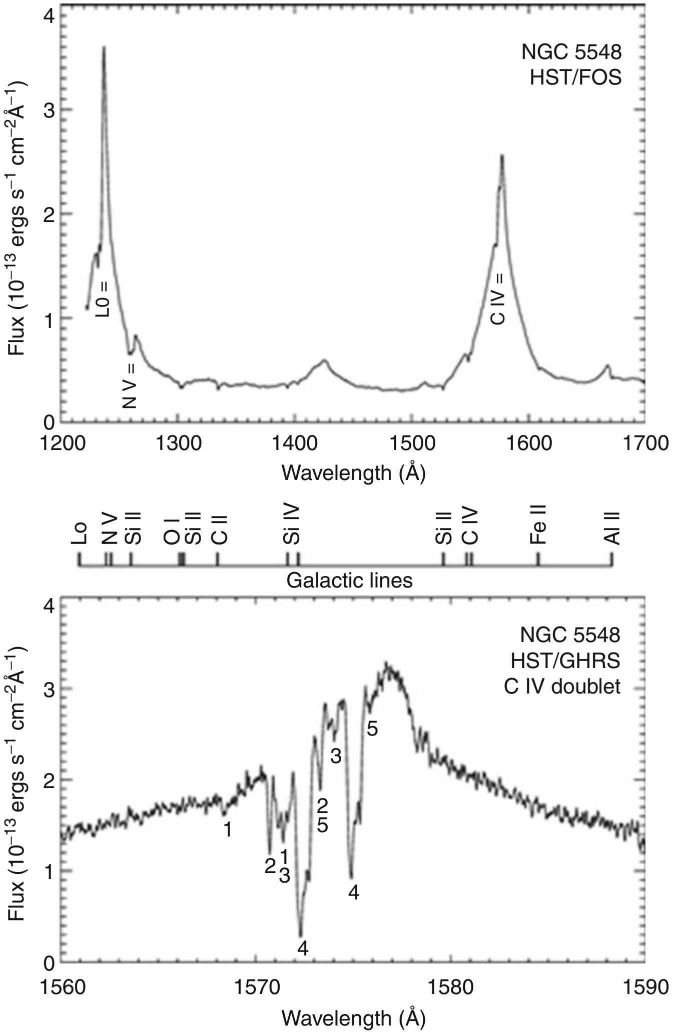


Fig. 7-37

HST spectra of the Seyfert 1 galaxy NGC 5548. The *top* panel shows a spectrum from the FOS and indicates positions of intrinsic UV absorption by H I Lα and the C IV and N V doublets. The *middle* panel shows the locations of interstellar absorption lines in the *top* panel due to our galaxy. The *bottom* panel shows a high-resolution spectrum in the C IV region from the GHRS and identifies five kinematic components of absorption in the C IV doublet (From Crenshaw et al. (2003, their Fig. 1))

[C IV] λ 1549 doublet shown in *Fig. 7-37*. In some cases, absorption seen in other lines is also seen including O VI λ 1034, Si IV λ 1400, Ly α, and Mg II λ 2800. High-resolution observations reveal that these features are often quite complicated, as with the complex shown in *Fig. 7-37*, which has several components. These widths are much wider than the expected thermal widths (a FWHM of 9 km s^{-1} would be expected for carbon at a temperature of 20,000 K). This indicates macroscopic motions within the absorbing gas.

A more spectacular form of absorption is seen in a minority of quasars, which exhibit broad features that extend from 1,000 to as much as 60,000 km s^{-1} blueward of the quasar's systemic velocity (Weymann et al. 1991; see examples in ➋ *Fig. 7-38*). These features, like their less spectacular Seyfert cousins, are most often found in UV lines, although they are sometimes seen in the Balmer lines. These features are typically highly saturated – in fact, so much flux can be removed from the rest-frame ultraviolet spectrum that some quasars with these features (called BAL, or broad absorption line, QSOs) will drop out of flux-limited optical samples. Therefore, there is significant debate about the exact fraction of quasars with these features: although flux-limited samples suggest the number is less than 10%, the extreme nature of some BAL features suggests that the number may be significantly higher, perhaps as high as one-sixth (Reichard et al. 2003a) of all quasars.

A wide range of BAL quasar types are seen – in some objects, known as miniBALs, the lines are relatively modest in width (only a few thousand km s^{-1} at most), while in others, the lines can be much broader. There is also a significant range in the types of lines where the BAL features are seen, with some objects showing them only in low-ionization lines (the LoBALs) and others showing them both in low- and high-ionization lines (HiBALs). The reader is referred to Reichard et al. (2003b) for the relative fraction of these subclasses, as drawn from the SDSS quasars. Delving into the taxonomy of each of these subclasses is beyond the scope of this work. However, their common feature is the overall spectral morphology, which shows a classic P Cygni-type line in the iconic objects, although in others where the absorption is heaviest and covers the largest velocity range, the amount of continuum and line emission that is absorbed makes the classical QSO spectral morphology of ➋ *Fig. 7-6* almost impossible to recognize.

Until very recently, BAL features were seen almost exclusively in radio-quiet objects. However, this is now known to be at some level a selection effect – BAL QSOs had first been identified in optically selected samples, and the complete radio-selected samples known at the time simply were not deep enough to contain these objects in significant numbers. This began to change with the identification of fainter radio quasars found in the FIRST sample (Brotherton et al. 1997, 1998) that were both clear BALs and clearly radio-loud. Later work with the SDSS verified that optically selected samples also contained these objects (insert comment about the fraction of these and whether there is any remaining dependence of BALnicity on radio-loudness).

The method that one would like to use to calculate the physical characteristics of these absorbing systems is similar to that used for the emission line regions: photoionization (➋ Sect. 5.4). This is the method of choice for the narrow absorption systems; however, in the more spectacular BAL QSOs, it is by no means clear that hydrodynamic ionization does not play a significant role.

Let us first consider the narrow absorption systems. One clue to their origin lies in the fact that in nearly all AGN with such features, the depth of at least one component is sufficiently large to indicate that the gas absorbs both the incident continuum as well as the broad emission line flux (as in the case of component 4 seen in the spectrum of NGC 5548, ➋ *Fig. 7-37*). This material must therefore lie outside of the broad emission line region. Calculations based on photoionization models typically yield density ranges $n \sim 10$–10^{10} cm^{-3} and ionization parameters of ~0.01–1 in the UV-absorbing gas; higher ionization parameters and densities are seen in the X-ray-absorbing gas. While these figures are somewhat lower than seen in the BLR, by itself this puts only loose bounds on the relative location of the absorbing gas, which from other considerations (see above) was already known to exist outside of the BLR.

A more recent constraint on this last question can be imposed from the existence of variability in these features, which is commonly observed on timescales of months to years

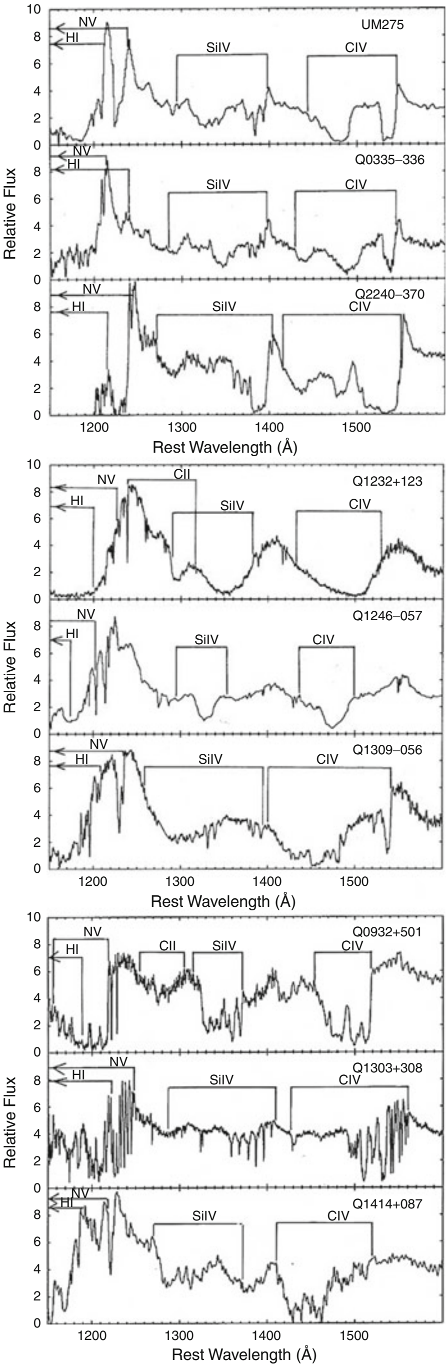


Fig. 7-38

Rest-frame ultraviolet spectra of several BAL quasars (Turnshek 1987, their Figs. 4–6). The features seen cover a considerable range in width and often consist of a complex of narrower subfeatures. In some objects, the features can completely destroy the classical QSO spectral morphology of *Fig. 7-6*

(e.g., Crenshaw et al. 1999, 2004; George et al. 1998; Kraemer et al. 2001a, b). Deep analysis of high-resolution, time-resolved spectra reveal that two mechanisms are at work: both changes in the ionization of the gas, due to variations in the ionizing continuum, as well as changes in the total column density due to, for example, bulk motion of the gas across the line of sight are required to explain the variability observed. Kraemer et al. (2001b) was able to map the absorbers in another Seyfert galaxy, NGC 4151, revealing a huge range in distances (from 0.03 to 2,150 pc from the nucleus, that is, extending from just outside the BLR to the larger galactic environment), number densities ($n \sim 10^{-3} \times 10^{9}$ cm^{-3}), and column densities ($N_H \sim 10^{18} - 3 \times 10^{21}$ cm^{-2}). Both number and column densities typically decrease with distance, but there are no other obvious correlations with distance.

BAL features are much more difficult to model. One reason for this lies in the morphology of the lines – whereas the narrow absorption features found in Seyfert spectra are almost entirely contained within the broad emission line's blue wing, this is not at all the case in BAL QSOs, which sometimes extend to blueshifted velocities much larger than the extent of the BLR. In addition, there is a much broader range of optical depths observed as a function of wavelength. Very often there are a significant range of wavelengths where the intensity goes to zero – that is, the broad absorption feature has removed essentially all flux from the spectrum. At other wavelengths, the absorption is only partial. The saturation of the absorption lines in many BAL quasars indicates that these flows are very high column densities – at least 10^{22} cm^{-2} and even higher, ranging up to 10^{24} cm^{-2}. Early X-ray observations also indicated that BAL quasars appear less luminous (Green and Mathur 1996), which more modern observations with *Chandra* and *XMM-Newton* suggest is due to the presence of large absorbing columns, which in several sources appear to be Compton-thick or nearly so (Brotherton et al. 2005; Shemmer et al. 2005; Gallagher et al. 2006; Giustini et al. 2008).

However, matters are not as simple as this. Naively, one would like to interpret the depth of absorption simply as an indicator of the optical depth and hence the absorbing column: where the flux goes to zero, the optical depth is several or higher, while where significant flux remains the optical depth is lower. However, the saturated nature of these systems means that particularly in the most optically thick systems it is very difficult to estimate the true column: the naïve calculation usually underestimates the true amount of material. This is partly due to the fact that the absorber does not have the same depth along all lines of sight to the continuum emission region (often referred to as *partial covering*). One might, for example, have very large optical depths along some rays and much less optical depth along others. The correct mean column density, averaged over the continuum source's projected area, is then seriously underestimated. Another complication is that higher-resolution observations of many BALs find that they have a considerable complexity in structure (➋ *Fig. 7-38*), similar to their narrower cousins. In this case, also, what appears to be partial absorption and modest optical depth could in fact be much larger optical depth at some velocities, but no absorption at others. The inference of partial covering is strongly supported by the observation that BAL quasars are much more highly polarized than non-BAL objects, with polarizations as high as 10%, concentrated in the absorption troughs (e.g., Ogle et al. 1999 and references therein). The simplest explanation of this is that the polarized light is scattered by dust in the BAL clouds by a mechanism similar to that invoked for Seyfert galaxies (e.g., electron scattering, as discussed by Wang et al. 2005). This also means that the polarized light was originally emitted along another line of sight than ours, and thus the column along that line of sight is different than our own.

Modeling of BAL systems indicates a higher ionization parameter than in narrow absorption systems, typically at least 1–10 given the presence of N V and O VI BAL features, but not too much higher, as Mg II, CIII] C IV are also seen. It is possible, if not likely, that a range of parameters is seen, but more work needs to be done here. It should also be noted that in BAL systems the ionization balance is also more complex, likely having contributions from both photoionization as well as hydrodynamic mechanisms. The distances of BAL regions from the central engine is not well constrained by ionization models. The best constraints come from observations of variability on timescales of months to years, which suggests distances of light-years or less from the nucleus. This is on the same scale as the BLR in quasars, as opposed to the narrow absorbers which are typically more distant. Another rough constraint can be derived from the fact that BAL features often absorb both line and continuum flux: this fact indicates that at least some of the BAL region lies beyond the BLR, although some overlap is possible.

The best estimates of the densities in BAL regions come from the combined use of ionization models with variability data, as pointed out by Krolik (1999). If, for example, the column density is observed to decrease along with the object's luminosity, then one may infer that recombination dominates ionization so that

$$n_e \approx 10^4 \left(\frac{\Delta t_{\rm rec}}{1\,{\rm year}}\right)^{-1} \left(\frac{Z_{\rm eff}}{4}\right)^{-2} {\rm cm}^{-3} \tag{7.39}$$

where $Z_{\rm eff}$ is the charge of the ion and $\Delta t_{\rm rec}$ is the timescale for an order unity change in the ionic abundance due to recombination. Alternatively, if the column density increases as the flux does, one would suggest that ionization by the increasing flux caused the change. Because the absorbing matter is exactly along the line of sight, it sees the same ionizing luminosity, so one can infer both an estimated location for the absorber (relative to the nucleus) as well as a density:

$$r \approx 500 \left(\frac{L_{\rm ion}}{10^{45}{\rm erg\ s^{-1}}}\right)^{1/2} \left(\frac{\Delta t_{\rm ion}}{1{\rm year}}\right)^{1/2} \left(\frac{Z_{\rm eff}}{4}\right)^{-2} {\rm pc} \tag{7.40}$$

$$n_H \approx 10^3 U^{-1} \left(\frac{T}{10^4\,{\rm K}}\right)^{-1} \left(\frac{\Delta t_{\rm ion}}{1\,{\rm year}}\right)^{-1} \left(\frac{Z_{\rm eff}}{4}\right)^{2} {\rm cm}^{-3} \tag{7.41}$$

where $\Delta t_{\rm ion}$ is the timescale for an order unity change in the ionic abundance (note that in this case photoionization dominates over any pressure terms). A third alternative is that the column might vary independently of the object's luminosity. In this last case, no such estimate is possible as the changes would be the result of the cloud moving across our line of sight. It should be noted that similar methods can also be used to infer the distance and density of narrow absorbing clouds.

The exact geometry associated with these absorbing clouds is uncertain. The fact that the narrow and broad variety of absorption features at least partially eat into the broad emission lines suggests that the material must be at a different orientation than the torus, discussed in the last section, as originally suggested by Weymann et al. (1991) for BAL QSOs. However, this is not at all clear, and more complicated geometries are certainly possible, particularly with a patchy torus composed of discrete, higher-density clouds as envisioned in the most recent models. What is, however, known is that these absorption features are evidence of large-scale mass outflows in AGN.

The kinematics observed in these lines show that the clouds that produce them are moving out from the nucleus with velocities of hundreds to tens of thousands of km s^{-1}, at distances of parsecs or more away from the nucleus. Such a flow will carry considerable momentum and will deposit large amounts of energy into the regions that surround the AGN, in the process also taking away from the active nucleus's immediate environs material that might be used to fuel future activity. The rate at which mass flows out is

$$\begin{aligned}\dot{M} &= 4\pi C_{\rm abs} r \mu_H \langle v \rangle N_H (r/\Delta r) \\ &= 0.04 \left(\frac{r}{\Delta r}\right) \left(\frac{C_{\rm abs}}{0.1}\right) \tau_{CIV} \left(\frac{X_{CIV}}{10^{-4}}\right)^{-1} \left(\frac{\langle v/c \rangle}{0.05}\right) \left(\frac{\Delta v/c}{0.1}\right) M_\odot \,\text{year}^{-1} \end{aligned} \tag{7.42}$$

where μ_H is the mean mass per H atom, r is the distance of the absorbing gas from the active nucleus, Δr is the radial thickness of the absorbing shell, $\langle v \rangle$ is the mean outflow velocity, ε is the efficiency of accretion, $\tau_{\rm CIV}$ is the optical depth in the C IV line, and $X_{\rm CIV}$ is the abundance of C IV. Particularly in luminous objects (e.g., quasars), this can be an interesting fraction of the accretion rate:

$$\frac{\dot{M} c^2 \varepsilon}{L} = 0.025 \left(\frac{r}{1\,\text{pc}}\right) \left(\frac{C_{\rm abs}}{0.1}\right) \tau_{\rm CIV} \left(\frac{X_{\rm CIV}}{10^{-4}}\right)^{-1} \left(\frac{\langle v/c \rangle}{0.05}\right) \left(\frac{\Delta v/c}{0.1}\right) \left(\frac{L}{10^{46}\,\text{erg s}^{-1}}\right)^{-1} . \tag{7.43}$$

Thus the existence of these clouds actually constitutes a major problem for BAL quasars as it can remove from the nuclear regions almost as much matter as is accreting into the black hole! Thus in these objects one is almost forced to conclude that the BAL stage must be transitory, a point discussed also later.

A second question one must ask ourselves before leaving the subject of these outflows is, exactly how are they driven? A number of mechanisms have been devised, including both thermal and radiation-pressure driven mechanisms. If one assumes that the gas is heated to a temperature such that its thermal energy matches its gravitational binding energy, then matter may be injected at some point; exposed to the AGN's radiation, its temperature rises, and once its temperature increases beyond the critical value, it is expelled. For this mechanism to work, a heating mechanism is needed. Many are possible, including photoionization, Comptonization, and collisions with energetic particles (see Begelman et al. 1991). If the heating rate is parameterized by that due to Comptonization, which may not be too bad an approximation as the rate of photoionization will be roughly dominated by radiative cooling, then the heating criterion is

$$\frac{L\sigma_T}{4\pi r^2} \frac{4k_B T_C}{m_e c^2} \frac{\phi}{\mu_e} > \frac{(\text{GM})^{3/2}}{2r^{5/2}} \tag{7.44}$$

The above assumes that gas is injected with the kinetic energy that corresponds to a circular orbit at radius r, μ_e is the mean mass per electron, and ϕ describes how different the true heating rate is from that derived from Comptonization. Once the wind takes off, its temperature may not need to maintain the Compton temperature T_C referred to above, because adiabatic expansion can compete with radiative heating. If one assumes that the wind achieves a Mach number of order 1, then the Compton heating rate is balanced by an effective cooling per (mean) unit mass $\sim (k_B T/\bar{m})^{3/2}\, r^{-1}$. That dictates a temperature of

$$\begin{aligned} T &\sim \frac{\bar{m}}{k_B \phi^{2/3}} \left(\frac{4k_B T_C}{m_e c^2}\right)^{2/3} \left(\frac{L\sigma_T}{4\pi r \mu_e}\right)^{2/3} \\ &\sim 5 \times 10^6 \left(\frac{T_C}{10^7 \text{K}}\right)^{2/3} \left(\frac{L}{10^{45}\,\text{erg s}^{-1}}\right)^{2/3} \left(\frac{r}{1\text{pc}}\right)^{-2/3} \text{K}. \end{aligned} \tag{7.45}$$

This assumption also allows us to estimate the maximum electron scattering depth as $\tau_{T,\max} \sim U^{-1} (v_{\rm orb}/c_s)^2 (L/L_E)$. In a thermally driven wind, the ratio of orbital speed to sound speed will be ~1, and because the critical ionization parameter for evaporation is ~10 (see the discussion in Chap. 10 of Krolik 1999), one can thus expect optical depths ~$0.1L/L_E$ whenever the luminosity is large enough to create the wind. These inferred conditions match fairly well to those found in the warm X-ray absorbers associated with Seyfert galaxies. If the terminal Mach numbers are few, then these winds can be expected to flow outward at several hundred km s^{-1} and develop clouds within the flow that are at temperatures of tens of thousands of degrees, which would match the properties of the narrow absorbers found in Seyfert 1 spectra.

Thus thermal mechanisms can drive the narrow absorption features but fail to achieve the velocities necessary to drive the winds in BAL quasars. In those objects, radiative driving must dominate. This mechanism is much more difficult to calculate, but if one assumes that the primary driving mechanism is due to Thomson scattering, photoionization codes may be used to compute the distribution of opacities $dN_l/d\kappa$ (U), that is, the number of lines N_l with center opacities κ. The equation of motion is then

$$v\frac{dv}{dr} = \frac{\mathrm{GM}}{r^2}\left[-1 + \frac{L}{L_E}M\left(N_H, U\right)\right]. \tag{7.46}$$

where $M(N_H, U)$ is a force multiplier that expresses the efficacy of line-driving over Thomson scattering. This parameter can be as large as 10–100 if the C IV line is marginally optically thick. It should be noted that radiative driving, augmented by winds, is a mechanism used to power the outflows in OB stars, so its presence in quasars should not be surprising. Calculations by Proga et al. (2000) and Chelouche and Netzer (2001) indicate that this mechanism can reach the velocities necessary to drive BAL outflows.

7 Jets and Lobes in Radio Loud AGN

The distinguishing feature of radio-loud AGN are their jets, as well as the hotspots and lobes seen at larger scale. These are features that, as already discussed, are also present in a significant number of radio-quiet AGN as well, albeit at much lower power levels.

7.1 Acceleration of Jets

A suite of both modeling and observational evidence indicates that these jets must originate deep within the central regions of an active nucleus. On the observational side, the evidence comes from high-frequency VLBI mapping of the M87 jet (➔ *Fig. 7-39*, Walker et al. 2009; Krichbaum et al. 2006), which shows that the jet extends very close to the central black hole and even reveals the location where collimation is taking place. The expansion of the jet's "cone" appears in these images on scales smaller than 0.5 milliarcsec, which corresponds to about 150 R_S for M87, and it continues in to distances of 0.1 milliarcsec from the VLBI core, which likely constitutes the base of the jet. Therefore, whatever process first collimates these relativistic outflows must occur on very small scales – within no more than 30 R_S from the black hole. Interestingly, however, the monitoring program that generated the images shown in ➔ *Fig. 7-39* did not show motions as fast as those seen on larger, arcsecond scales (➔ *Fig. 7-22*), where speeds as high as ~6c have been seen. On milliarcsecond scales, it is necessary to image the jet every few weeks to reveal relativistic motions, but indeed, speeds as high as 2c are seen in the monitoring data. On the smallest scales (within 0.5 milliarcsec), the *counterjet* – that is, the flow that

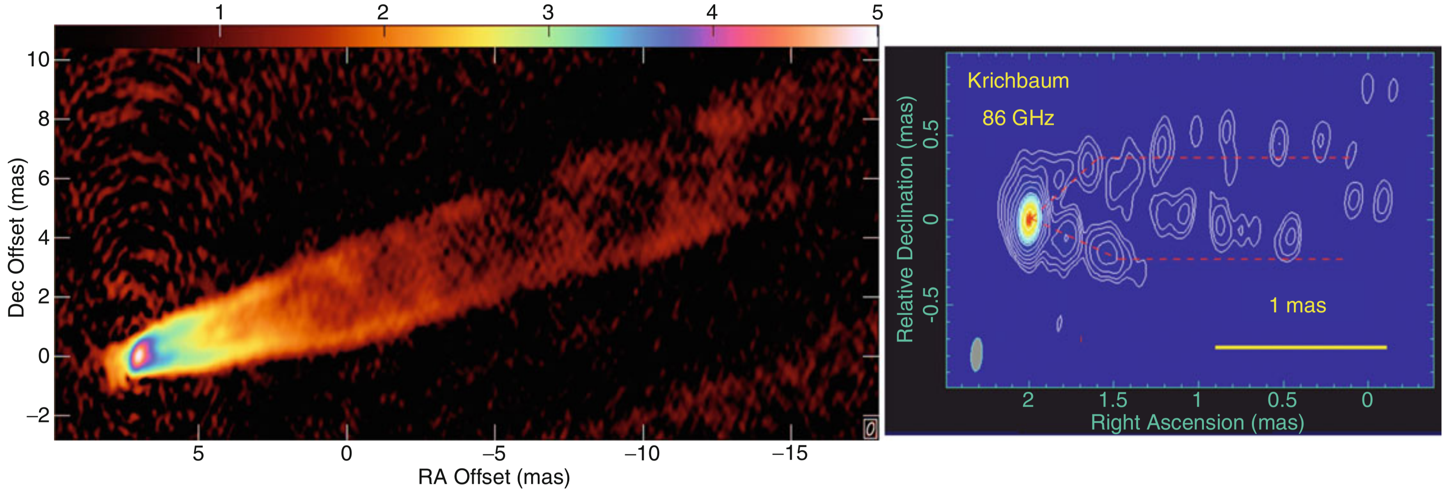


Fig. 7-39
High-frequency VLBA mapping of the M87 jet, at (at *left*) 43 GHz and (at *right*) 86 GHz. These images show the very innermost workings of a radio jet. Importantly, note that the jet takes up a constant angle up until the very smallest scales, when the 86-GHz image reveals that the initial expansion and collimation region can be seen (traced out by *red dotted lines*) within the innermost 0.5 milliarcsecond of the jet. For reference, at the distance of M87 (16 Mpc), 1 milliarcsecond is 0.078 pc = 16,000 AU, or 300 times the Schwarzschild radius for a 3 × 10^9-$M_\odot$ black hole (Figures from Walker et al. (2009) and Krichbaum et al. (2006))

emerges in the direction opposite to the jet seen at arcsecond scales – is measured. One can also see in ➋ *Fig. 7-32* that the images of the jet have a very edge-brightened appearance. This suggests that they are located along the jet's outer edge, often called the "sheath." Only much fainter components can be seen in the inner part of the jet's cross section, the "spine." Many jet models (e.g., Ghisellini et al. 2005) postulate that the spine should move at faster speeds, similar to the center of the flow of a fire hose. There are no detected motions in this central region, but the data are not deep enough to say this with confidence. However, it is clear that the *acceleration* process must take place over a significantly larger scale than does the *collimation* of jets.

This recent observation set constitutes a challenging goal for models of jet generation to meet – they need to be able to simulate the region very close to the central black hole, certainly within a few tens of gravitational radii if not all the way to the ergosphere. Thus any model for jet generation must include both special and general relativity, as well as a variety of other physics, most particularly magnetohydrodynamics (MHD). Moreover, if it is to model the generation, acceleration, and collimation of the jet, it needs to cover a large dynamic range in angular scale. This makes it very challenging for computational modeling. Because of this complexity, models of jet generation typically make a number of simplifying assumptions, in particular that of axisymmetry, as well as essentially infinite conductivity, which together allow the equations of ideal, relativistic MHD to be cast in a semianalytic form (e.g., Li et al. 1992; Contopoulos 1994; Meier et al. 2001; Vlahakis and Königl 2003, 2004) via self-similarity arguments.

It is beyond the scope of this chapter to give a full treatment of the equations of MHD. However, the differences between regular fluid flow and ideal MHD flow can be understood by referring to ➋ *Fig. 7-38* (left-hand panel), which shows a three-dimensional GR-MHD simulation of a magnetized jet's propagation. As shown by ➋ *Fig. 7-40*, the magnetic field lines thread the plasma and are frozen in. These field lines restrict the flow of plasma to the parallel direction. Moreover, if the field is strong (i.e., if the plasma's hydrodynamic pressure ρv^2 is less than the magnetic pressure $B^2/8\pi$) and anchored in a rotating star or disk, then any plasma trapped in the field will be flung centrifugally outward along the field lines. Another important point is that parallel magnetic field lines tend to repel each other. This produces a pressure on the plasma perpendicular to the field lines due solely to the magnetic field. A weak field can thus be strengthened by bringing together many weak parallel lines of force to produce the equivalent of a few strong ones. As a result, compression perpendicular to the field lines or toroidal coiling can enhance the field. Finally, magnetic field lines do not maintain a curved shape unless they are acted on by forces from the plasma or other field lines – if left alone, they will straighten like springy wires, and if coiled in a hoop or spiral, the field will try to shrink around its axis to eliminate all but the straight axial component.

Simulations (➋ *Fig. 7-40*, center and right panels) begin with a disk that is initially in rotational equilibrium about the central black hole. Such a disk naturally rotates differentially. The disk is also threaded with an axial (vertical) magnetic field that is sufficiently strong to exert a braking force on the rotating plasma. As the simulation progresses, the magnetic field's braking force removes angular momentum from the torus, transferring it up along magnetic field lines into the coronal plasma, which is also frozen to the field lines. As these rotating magnetic twists propagate out, they push out and pinch the coronal plasma into a spinning jet. As the disk material loses angular momentum, it falls toward the central object, releasing half of its gravitational energy into kinetic energy of rotation that continues to power the outflow. Thus the production of a high-velocity outflow or jet is a natural and mandatory byproduct of the accretion process where a strong magnetic field is present. The same process that generates the jet is also behind the magnetorotational instability, and therefore it may be that the presence of

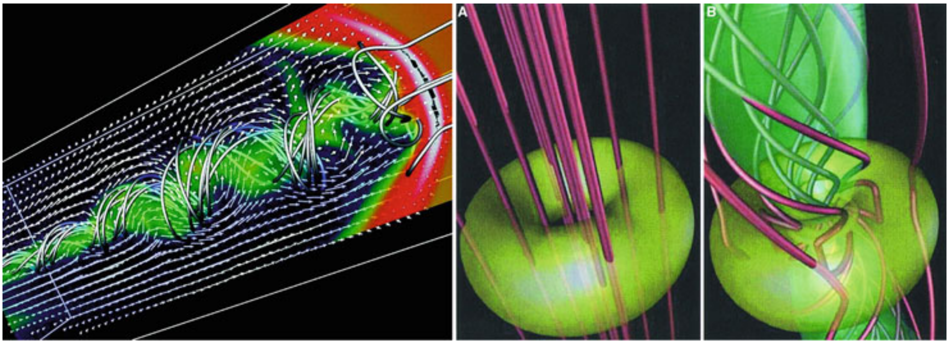


Fig. 7-40

Propagation (*left*) and generation (*center* and *right*) of a magnetized jet by a supermassive *black hole* and its accretion flow. The diagram at *left* shows flow velocity (*arrows*), plasma density (colors, with *white* and *blue* lines indicating high and low pressure, respectively), and magnetic lines of force (metallic tubes). In this simulation, the initially axisymmetric, rotating jet has developed a helical-kink instability that distorts its shape; such an instability may explain some of the wiggles observed in parsec-scale jets. The feature at far right in this diagram is a strong shock wave generated as the super-Alfvenic flow propagates into a region with decreasing Alfven velocity. The diagrams at *center* and *right* were generated from a simulation that included a thick, magnetized toroidal disk surrounding a central 10^8 $M_{\odot}$ *black hole*. The *center* panel shows the initial state, with the disk in rotational equilibrium with the axial magnetic field. As time goes on, the differentially rotating disk drags the field lines in the azimuthal direction, creating a braking force that allows material to accrete inward and gain additional rotational energy. This produces a torque on the magnetic field and generates a spinning plasma jet that carries away matter, angular momentum, and energy, producing the configuration seen at right (Figures from Meier et al. (2001, their Figs. 2 and 5))

such an axial magnetic field and hence the jet is a critical factor in removing angular momentum from disk material as it spirals inward toward the black hole.

Most simulations show that the speed of the ejected outflow is close to the escape speed at that radius – in other words, the magnetic field lines that are anchored to the disk closer in to the black hole will produce a faster flow than those that are anchored at greater distances. Thus these simulations reproduce the fire-hose-like velocity structure ("spine-sheath," see above) hinted at by the high-frequency VLBA images of M87. Simulations have been done both for nonrotating (Schwarzschild) black holes as well as strongly rotating (Kerr) black holes. An outflow is generated in both cases, with a speed that is directly related to the escape velocity at the location of the innermost stable orbit, which represents the accretion disk's inner edge. In the Schwarzschild case, the innermost stable orbit is at 3 R_S, which makes the jet production region slightly larger, and hence the speed of the outflow is ~0.5c. However, in the maximal Kerr case, stable orbits can exist right down to the event horizon, making the center of the jet production region much smaller (similar in size to the ergosphere) and generating a faster outflow (~0.9c, i.e., Lorentz factor $\Gamma = 2$). Thus the production of a relativistic jet is most likely tied to the spin of the black hole. Furthermore, even if the disk surrounding a Kerr black hole is initially given no rotation, a jet is still formed. This is because as matter plunges in to the ergosphere, it becomes caught up in the rotating space around the black hole and then routed out into the jet. Thus the production of the jet is intimately rooted in the coupling of the magnetic field with the black hole's spin – the so-called Blandford–Znajek (1977) mechanism.

It should be mentioned that the generation of jets through a thin accretion disk produces a rather small luminosity. Therefore most jet generation models require a thick, more toroidally shaped disk. This can be produced by advection-dominated accretion flows (ADAFs), where most of the disk's thermal energy is carried into the black hole, as well as convection-dominated solutions. These increase the power that one can generate through the Blandford–Znajek mechanism by orders of magnitude. However, to produce the most luminous quasars, it appears necessary to combine the Blandford–Znajek mechanism operating on a thin disk with a rapidly rotating Kerr black hole. This will have the effect of naturally thickening the disk on small scales, through the frame dragging near the black hole's ergosphere.

One additional factor must be considered, namely, the fact that the Blandford–Znajek mechanism does not by itself appear to accelerate jets to speeds fast enough to account for those seen in VLBI monitoring (up to 40c), nor can it account for their collimation. To get Lorentz factors up to 40 requires an additional, distributed acceleration process. This is provided by ideal MHD acceleration, which is by itself quite efficient, and powered by the pressure gradient in the toroidal magnetic field within the rotating jet. It can be thought of as akin to the gradual unwinding of a twisted rubber band and is by its very nature spatially extended. It naturally produces a rough balance between the Poynting and kinetic fluxes. The process develops naturally with the jet core due to compression by the hoop stress in the toroidal magnetic field. It is this same process that helps to confine the inner core of the magnetic field; however, the outer sheath of the jet must be confined by external pressure. These two can be driven together by a disk wind, which might develop hydrodynamically and are also seen in radio-quiet AGN.

Another issue that is not well constrained is the makeup of jets – whether they are positron/electron pair plasma, hadron dominated, or carry most of their energy as Poynting flux. Part of the problem is that there are no easy diagnostics given the nearly complete ionization of the jet material. Various methods have been derived to answer this issue, most of which have so far come up with a negative answer, both at large and small scales, most of which have to do with using the jet material as a Comptonization target (see ➋ Sect. 7.3) and then predicting the response. For examples, see Georganopoulos et al. (2005), Begelman and Sikora (1987), and Sikora et al. (1997).

7.2 Propagation and Dynamics of Jets at Larger Scales

Once launched, the jet can carry away a major fraction of the kinetic power and angular momentum that was within the disk. A jet of high power thus will remain collimated for very long distance, flowing through the galaxy and out beyond it into the surrounding cluster – as is indeed seen in virtually all sources. As the jet propagates, the magnetic field is carried along with it, and once the jet has reached its terminal velocity, the magnetic field lines will be essentially along the jet direction.

Jets are prone to a number of instabilities. Included among these are Kelvin-Helmholtz instabilities, as well as waves, pinches, and bends. Simulations of these have been performed by many workers, including Aloy et al. (2003), Hughes et al. (2002, 1996), Hardee et al. (2001, 2005, 2007), Hardee (2003), and Hardee and Rosen (2002). These simulations show that instabilities, perhaps combined with natural variations in the flow of material down the jet, can produce many of the moving and stationary features in VLBI and VLA maps and also provide a natural mechanism for producing flares such as those seen in blazars as well as the M87 jet. As instabilities propagate, shock waves are generated, changing the direction of the local magnetic field.

Polarization observations, both in the radio and in the optical, reveal the complexity of many of these features. An example is shown in *Fig. 7-41*, which shows VLA and HST polarimetry of the jet of M87 (Perlman et al. 1999) on arcsecond (kiloparsec) scales. These maps reveal a wealth of detail, with complex structure in several of the knots, including evidence for compressed magnetic fields (as predicted in shock models) as well as rotations in the magnetic field. Several jet regions show a higher polarization at the jet edges, indicating shear by the surrounding interstellar medium. At least four inner jet knots show strong decreases of polarization at the position of the flux maximum. Closer examination of these figures shows that several of the knots have changes in optical polarization that are not seen in the radio. These include large

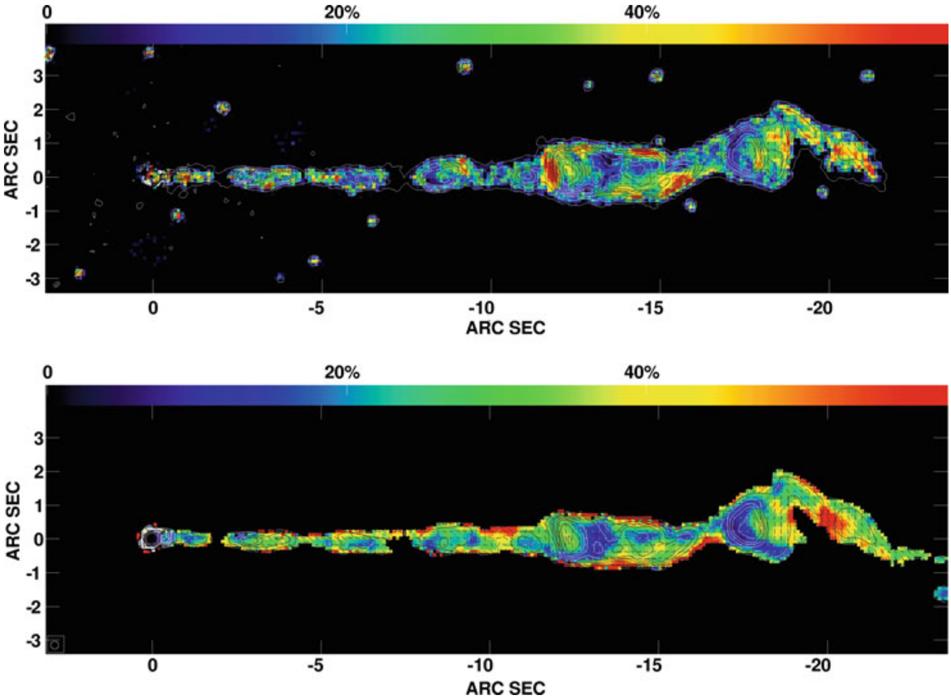


Fig. 7-41

Imaging polarimetry of the M87 jet, as obtained with the HST (*top*) and VLA (*bottom*). The *colors* indicate fractional polarization in each band, while the contours represent total flux in the optical and radio, respectively. A polarization of 20% is typical of what we see in the optical, with somewhat higher polarizations seen in the radio. The flux maximum regions of several knots exhibit large changes in the optical, including perpendicular magnetic fields and reduced polarization, that are not seen in the radio. These suggest that the optical emission from the M87 jet comes mainly from a region distinct from the radio emission (suggested to be the jet spine), while radio emission comes from the entire cross section (Figures adapted from Perlman et al. (1999))

rotations (near 90°) in magnetic field direction as well as decreases in the polarization fraction, in some cases to near zero. The explanation suggested for this phenomenon was that the optical emission and the jet shocks come largely from the jet's interior or spine region, while the radio emission comes from the entire jet cross section. Shocks would accelerate particles and thus be seen as enhanced emission first within the optical. Later multiband imaging (Perlman et al. 2001a) supported this scenario, as dramatic spectral index changes are seen in knot regions in the optical, but not in the radio. Later work on several other FR I radio galaxy jets reveals a similar pattern – not only do the jet dynamics reflect themselves in the polarimetry in the optical but also a spine-sheath pattern to the jet flow appears common, with more subtle changes also seen that indicate a variety of different disturbances in those jets, including twists as well as other features.

Emission at higher energies (optical, X-rays) is seen in over three dozen jets now, thanks to observations by the HST and Chandra (e.g., ➋ *Fig. 7-42*). The next section discusses the emission mechanisms for these large-scale jets in the next section; however, it is worth noting here that optical and X-ray emission from jets is often concentrated within the brighter knots in the inner jet and often does not appear to occupy the entire jet cross section.

Finally, it is worth discussing in this respect the effect that the jet will have on its surroundings. As the jet propagates, it carries with it an enormous amount of kinetic flux (up to 10^{61} ergs), and even though the flow is usually lower density than the surrounding medium, because of its high speed, it will deposit large amounts of energy into the surrounding regions. A number of effects can be stimulated, including star formation as well as hydrodynamic shockwaves that result in heating of the galactic and cluster medium. Simulations by O'Neill et al. (2005) indicate

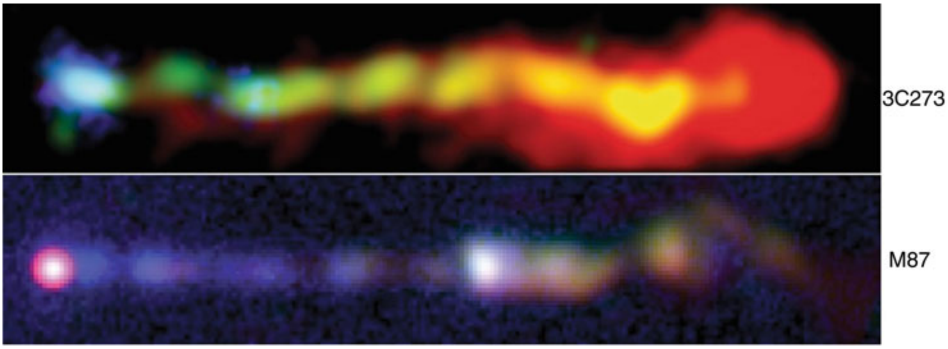


Fig. 7-42
Multispectral appearances of the jets of M87 and 3C273. The optically visible portion of the jet of M87 (shown) extends about 7,000 light-years from the nucleus (at *left* in the image at *bottom*) and feeds plasma to lobes that span 200,000 light-years. The bright portion of the jet of 3C273 extends for a projected distance of 90,000 light-years, beginning with a bright knot approximately 100,000 light-years from the host galaxy's nucleus (which is 12′′ off to the left side in the image shown). These figures were made using data from the radio VLA (*red*), infrared Spitzer space telescope (*yellow*), Hubble (*green*), and the Chandra X-ray observatory (*blue*). In both panels, *red colors* represent regions that are brightest in the infrared and/or radio, whereas *blue* colors represent regions that are brightest in the X-rays. Note that in both jets, the X-ray and optically bright regions of the jet do not occupy the entire cross section

that even light jets will deposit approximately half of their kinetic flux as thermal energy in the surrounding medium. Half of this would go directly into dissipative heating of the ICM, which is believed to be needed to support the reheating of the cluster gas against cooling flows. The remainder of the energy would reside primarily in the surrounding cocoon, which has been observed in jets on large scales (see ➋ *Fig. 7-8* for examples). Interestingly, these same simulations show that jets at low Mach numbers are more easily bent and disrupted. As a result, it seems likely that the Fanaroff-Riley divide is not only a reflection of the power of the outflow and AGN process but also the Mach number of the jet flow.

7.3 Emission Mechanisms in Jets

The emission from the jets and lobes of active galaxies at lower (radio through optical) energies is normally interpreted in the framework of *synchrotron radiation* from relativistic electrons whirling around in the magnetic field of the jet. Basic electromagnetic theory shows that electrons in a magnetic field propagate along helical, that is, corkscrew-shaped paths, so that they are constantly accelerated by the Lorentz force. Accelerated electrons also emit electromagnetic radiation in the form of synchrotron emission. If an electron has Lorentz factor γ, this emission has a characteristic frequency

$$\nu_c = \frac{3\gamma^2 eB}{4\pi m_e c} \approx 4.2 \times 10^6 \gamma^2 \left(\frac{B}{1\mathrm{G}}\right) \mathrm{Hz}, \tag{7.47}$$

where B is the magnetic field strength, e is the electron charge, and m_e is the mass of the electron. For frequencies considerably lower than ν_c, the spectrum of a single electron is proportional to $\nu^{1/3}$, whereas at larger frequencies it decreases exponentially. As the synchrotron emission is seen over several decades of frequency, to a first approximation, the spectrum of a single electron can be considered as quasi-monochromatic, as it radiates over a very small energy range. To produce radiation at centimeter wavelengths in a ~100 μG magnetic field, Lorentz factors $\gamma \sim 10^4$ are required, that is, the electrons need to be highly relativistic. To attain such high energies, particles must be accelerated very efficiently in the inner regions of AGN, particularly in some objects where synchrotron radiation is observed up through X-ray energies. The mechanism for this particle acceleration is believed to be diffusive shock acceleration. Diffusive shock acceleration naturally produces a power-law distribution of particle energies (see e.g., Longair 1994), with a number density $N(E)\,dE \propto E^{-s}dE$. It is easy to show that the synchrotron radiation from a population of particles that has a power-law energy distribution will also be a power law, with spectral index $\alpha = (s-1)/2$. An observed spectral index of $\alpha = 0.7$ results from a particle energy spectral index of $s = 2.4$, which is very similar to the energy distribution of cosmic rays observed from Earth. Thus the synchrotron emission observed in AGN results from a population of energetic electrons that covers a wide range of energies.

At low frequencies, the optical depth for absorption due to the synchrotron process can become significant. When this optical depth is close to or larger than 1, a source is called synchrotron self-absorbed. Such a source can have a significantly flatter spectrum, and for a small frequency interval, it may even rise. In the limiting case of infinite optical depth, the spectral shape approaches $S_\nu \propto \nu^{2.5}$ at low frequencies. It is synchrotron self-absorption that accounts for the rather different spectral morphology of the compact radio-emitting components.

The electrons in the emitting region lose energy through emission. The power emitted by an electron of Lorentz factor γ, integrated over all frequencies, is

$$P = \frac{dE}{dt} = \frac{4}{9}\frac{e^4 B^2 \gamma^2}{m_e^2 c^3} = \frac{4}{3}\sigma_T c \beta^2 \gamma^2 U_B \tag{7.48}$$

where, in the last form, U_B is the energy in the magnetic field. The characteristic time in which an electron loses energy is then obtained from its energy $E = \gamma m_e c^2$ and its energy loss rate $\dot{E} = -P$ as

$$t_{\rm cool} = \frac{E}{P} = 2.4 \times 10^5 \left(\frac{\gamma}{10^4}\right)^{-1} \left(\frac{B}{10^{-4}\rm G}\right)^{-2} \text{ years.} \tag{7.49}$$

As the cooling time $t_{\rm cool}$ depends on γ^{-1}, it is thus dependent also on ν^{-1}. For low-frequency (meter-wave and longer) radio emission, this lifetime is longer than or comparable to the age of radio sources. But at higher frequencies, the cooling time becomes important, with optical synchrotron-emitting particles having radiative lifetimes typically in the hundreds of years, and X-ray synchrotron-emitting particles having radiative lifetimes of only a few years. The fact that optical and X-ray synchrotron emission are observed in a significant number of objects (see later in this chapter) means that the processes of particle acceleration are not confined to the innermost regions of an AGN but must also occur at large distances (many kiloparsecs) from the central black hole.

Since the characteristic frequency (➋ 7.47) of synchrotron emission depends on both the radiating particle's Lorentz factor γ and the magnetic field B, it is impossible to measure both quantities independently. In order to estimate the magnetic field strength, it is often assumed that energy is divided equally between the particles and the magnetic field (this is often called *equipartition*). A second, more sophisticated approach is to first estimate the magnetic field such that the total energy of the relativistic electrons and magnetic field is minimized for a given luminosity. It can thus be shown that equipartition represents the minimum energy state for the macroscopic system and is thus a plausible, if not a physically accurate, representation of the physical state of the radiating region.

A second process, that of inverse-Compton scattering (already discussed in ➋ Sect. 4.3), also occurs in AGN jets. It is trivial to show that if the input spectrum is a power law, then the inverse-Compton process produces a power law with the same slope. The photons being scattered (the so-called *seed photons*) can come from a variety of sources – either from the jet itself (often referred to as synchrotron self-Compton), the broad or narrow-line regions, the torus, starlight, or even the cosmic microwave background. Models of AGN jet emission must take all of these regions into account as possible seed photon sources. In ➋ *Fig. 7-43*, an example jet spectrum is shown that takes into account not only the synchrotron process but also Comptonization from broad-line photons.

Two issues must be mentioned when discussing ➋ *Fig. 7-43*. The first is that multiple Comptonization processes are possible. One uses synchrotron photons generated within the jet (called synchrotron self-Compton or SSC). For SSC, so long as one can use the Thomson cross section (see below), one can derive that the power radiated via the inverse-Compton process is

$$P = \frac{4}{3}\sigma_T c \beta^2 \gamma^2 U_{\rm rad.} \tag{7.50}$$

This is almost identical to the form discussed earlier (➋ 7.49) for synchrotron energy, with the exception that in place of U_B, the energy density in the magnetic field, one has $U_{\rm rad}$, the

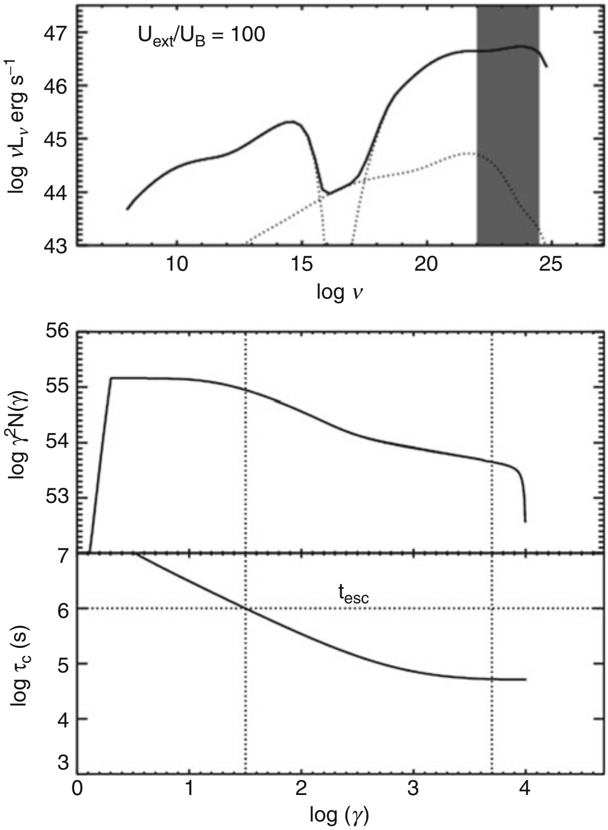


Fig. 7-43

A model for an external Compton-dominated blazar. Here the assumed ratio $U_{\mathrm{ext}}/U_B = 100$ and we have assumed a source size ∼5×10^{16} cm to ensure the dominance of EC over SSC. *Bottom panel*: the electron cooling time as a function of γ. *Middle panel*: the electron energy distibution. *Top power*: the emitted power. *Solid line* represents total power, and *dotted lines* represent synchrotron (*left-most*), SSC (*central*), and EC (rightmost and most powerful) components. The *gray band* is roughly the EGRET-GLAST regime (Figure taken from Perlman et al. (2008, their Fig. 6))

energy density in radiation. Thus the competition between these two processes in terms of the dominance of energy loss mechanisms breaks down to the ratio

$$\frac{P_{\mathrm{sync}}}{P_{\mathrm{IC}}} = \frac{U_B}{U_{\mathrm{rad}}} \tag{7.51}$$

At most energies, the cross section for Compton scattering is simply the Thomson cross-section. However, at high energies, when $x = h\nu/m_e c^2 \geq 1$, it is necessary to use a result from quantum electrodynamics, namely, the Klein-Nishina cross section,

$$\sigma_{\mathrm{KN}} = \frac{3}{4}\left\{\frac{1+x}{x^3}\left[\frac{2x\,(x+1)}{1+2x} - \ln\,(2x+1)\right] + \frac{1}{2x}\ln\,(2x+1) - \frac{1+3x}{(2x+1)^3}\right\}. \tag{7.52}$$

The Klein-Nishina cross section is an energy-dependent modification to the Thomson cross section. At low energies ($x \ll 1$), it reduces to the Thomson cross section, but at high energies it decreases roughly as $1/x$. A rough rule of thumb is that the Klein-Nishina cross section is dominant when $\varepsilon_o \gamma \sim 1$ where ε_o is the photon energy and γ is the Lorentz factor of the particles.

A second type of Comptonization process is also possible, where the seed photons come from outside the jet. This process, called external Compton or EC, can use seed photons either from the BLR (at small scales), the torus, starlight, or (at the largest scales) the cosmic microwave background. Which process will be dominant in a given jet depends only on the energy density of the various photon populations in the jet's comoving frame.

Thus, it is noted that ➲ *Fig. 7-43* was created with parameters that match 3C 279, a typical Compton-dominated source. One can see in ➲ *Fig. 7-43* a number of features due to the Klein-Nishina cross section. For example, there is a "hump" at optical-UV energies, where both jet and BLR seed photons are present with $\varepsilon_{\mathrm{w}} \gamma \sim 1$. As a result, the scatters that produce GeV photons will occur between the Thomson and Klein-Nishina regime, producing a flat or rising GeV spectrum, as well as achromatic GeV variability, with the latter occurring because electron cooling is not energy dependent in the valley between the Thomson and Klein-Nishina regime. These predictions are just beginning to be tested with data. CGRO observations of 3C 279 find a fairly flat spectrum at high energies (Joshi et al. 2011; For a counterexample, however, see Ackerman et al. 2010).

The second remark that needs to be made concerns the nature of the variability one will observe in a given source, for it is through observing variability in coordinated, multiwaveband campaigns that models of this sort can be tested. In most of the environments within AGN jets and lobes, the source is optically thin – that is, a given photon will scatter at most once before it escapes the source. Therefore the response that one would expect at gamma-ray energies would be what is known as quadratic variability – that is to say, if the source increases in flux by a factor of 2 at lower energies (i.e., in its synchrotron component), an increase of a factor 4 in its gamma-ray flux would be seen. This is the expectation if a source's inverse-Compton emission is roughly equal or less than its synchrotron emission, as in BL Lacs. This is shown in ➲ *Fig. 7-43* by the lowest-luminosity set of curves.

This argument led early workers, notably Wehrle et al. (1998), to argue that due to the observation of superquadratic variability in 3C 279, an additional seed photon source, believed to be the BLR, might be needed. In 3C 279, a bright blazar with prominent broad lines, it might be natural to expect EC to contribute significantly to the observed high-energy emissions of the source; it is important to note that the data do not yet require BLR seed photons to be invoked. However, while attractive, this argument is not strictly correct.

In more luminous sources, such as 3C 279, the efficiency of Comptonization increases, as the increased number of seed photons increases the number of photon-electron collisions. After a certain point, the source will not be optically thin to Comptonization, and in fact in the most luminous sources it is Comptonization, rather than synchrotron emission, that represents the primary cooling channel for jet electrons. In these sources, the jet becomes optically thick to Compton scattering. As a result, one can get a superquadratic variability response, as pointed out by Georganopoulos et al. (2006), even without invoking an external seed photon source. ➲ *Figure 7-44* shows how the response of the spectrum changes as the luminosity of the synchrotron component, and corresponding dominance of the Compton channel goes up.

Observational data are now becoming good enough where direct tests of some of these models are possible. The most direct way to test these models directly is with multiwaveband campaigns, which use the fact that different physical models of emission and/or jet

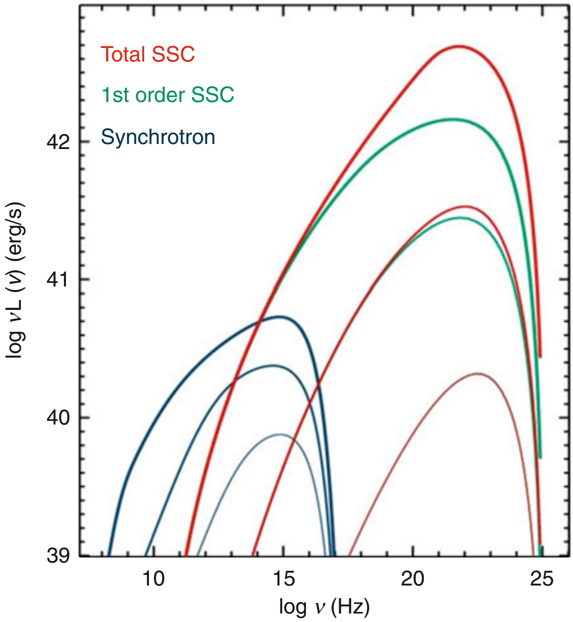


Fig. 7-44

Example spectral response for variable sources. Synchrotron emission is seen in the *left-hand* curves, whereas SSC emission is seen in the *right-hand* curves. As can be seen, as the dominance of Compton cooling increases, the source becomes optically thick to Comptonization. These second-order reactions cause superquadratic variability without the need for an external seed photon source

compression/expansion predict different responses from different jet bands. If for example, one is observing purely synchrotron emission where the main jet cooling process at X-rays is due to synchrotron emission, one would expect to observe not only correlated variability between bands but also an energy-dependent response during the cooling phase, with less energetic photons taking longer to cool, and the jet spectrum softening during the cooling process. This has been observed in the X-ray band during many BL Lac observations (e.g., Fossati et al. 2008). An alternate possibility, however, is that adiabatic expansion may dominate the cooling process, at least between bands, and this would predict no change in the SED as the source cools.

Multiwaveband campaigns can be used to test the latter set of comments regarding superquadratic variability. This was done recently by Aharonian et al. (2009) for the exceptional gamma-ray variability episodes of PKS 2155–304 during 2006. As can be seen (➲ *Fig. 7-45*), at that epoch, PKS 2155 exhibited variations that were very similar in X-ray and TeV gamma-rays, but those bands were not correlated with the optical emission. Moreover, the response of the SED to the variations was not the classical quadratic one would expect – instead, nearly cubic behavior was commonly seen, as shown in ➲ *Fig. 7-38*. This demonstrates the viability of the SSC mechanism for producing superquadratic variations. However, the lack of correlation between the X-ray/TeV and optical lightcurves led Aharonian et al. (2009) to invoke a multizone model.

As noted earlier, high-energy emissions are also seen from jets on kiloparsec scales. Here it is not so easy to find diagnostics to test the viability of emission mechanisms. While at optical

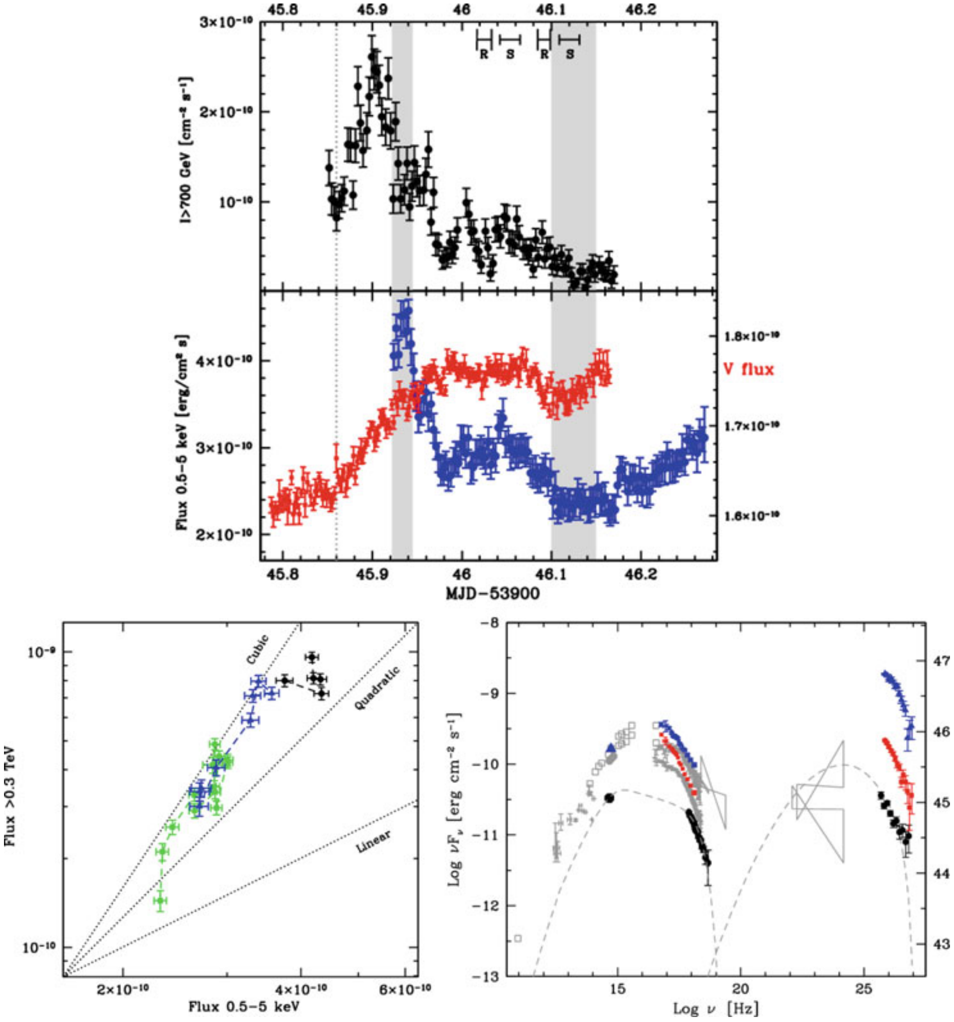


Fig. 7-45

Multiwavelength observations of PKS 2155–304 during 2006. At *left*, the optical (*red*), X-ray (*blue*), and TeV (*black*) lightcurves observed during July 29–30, 2006. Note the correlated X-ray/TeV variability, but the lack of correlation seen between those bands and the optical emission. At *center*, the spectral energy distributions are plotted, showing respectively the highest and lowest simultaneous states during this night, together with historical data (the latter in *gray*). The highest state is represented by the *blue* symbols, while the lowest state is represented by the *red*. The *black* points and accompanying curves represent data from the 2003 multiwaveband campaign with a 1-zone fit (Figures from Aharonian et al. (2009, respectively Figs. 2, 15, and 13))

and radio energies, synchrotron emission is the natural interpretation, supported by the spectral shape and high polarization; at X-ray energies synchrotron, and Comptonization mechanisms are both possible. For FR I sources, it is likely that synchrotron emission is the dominant X-ray emission mechanism. This is for several reasons. First of all, the observed optical-radio emission component is seen to extrapolate to X-ray energies, albeit with breaks that may be steeper than the classical 0.5 and evidence that only a fraction of the jet cross section radiates at these high energies (see, e.g., Hardcastle et al. 2001; Marshall et al. 2002; Perlman and Wilson 2005). In addition, the variability observed in M87's knot HST-1 weighs heavily in favor of synchrotron X-ray emission.

For more luminous, FR II sources, however, the jury is still out on the X-ray emission process (Harris and Krawczynski 2002, 2006; Hardcastle et al. 2004). The SSC mechanism is highly unlikely because it requires the jet to be massively out of equipartition; moreover, the observed morphology, whereby the jet emission is seen primarily within knots, also argues against it. However, two other mechanisms are possible, namely, synchrotron emission and Comptonization of the cosmic microwave background photons. Both processes appear to be viable at present. In some sources, for example, 3C 273 (shown in ➋ *Fig. 7-40*) and PKS 1136–135, the optical-UV spectrum appears to extend to high energies and connect with the Chandra emissions, and high optical polarizations are detected (Cara et al. 2010), the latter ruling out Comptonization of an unpolarized photon population (Uchiyama and Coppi 2010; McNamara et al. 2009). However, in other sources, notably PKS 0637–752 (Mehta et al. 2009), the valley in between the radio-optical and X-ray components is extreme, and a separate mechanism is needed, which is more likely to be EC/CMB.

8 Final Remarks

This chapter has discussed active galactic nuclei in considerable depth, including their properties, the current state of unified schemes, and the emission mechanisms required. As detailed in the discussion above, the evidence for the unified scheme is quite strong, so that active galaxies comprise a particular subset of the supermassive black holes that lie at the center of all bright, massive galaxies. As detailed in further chapters (➋ Chaps. 10 and ➋ 11), these galaxies – and thus the black holes located at their centers – presumably grew hierarchically from initial density perturbations through merging of smaller galaxies into the objects seen today.

Yet despite all this discussion, the one question left unexplored is fairly basic – namely, what actually makes a given galaxy's nucleus become active. It is fairly easy to prove that 100 million solar mass black hole, accreting at the Eddington rate, will exhaust the nuclear medium that feeds its activity in a small fraction of the Hubble time, ~a few tens of millions of years. Thus it is important to realize that the AGN phase represents a small fraction of the life cycle of a given galaxy. Moreover, in order to maintain the number of AGN seen today, it is necessary to assume that most or all bright galaxies go through an AGN phase at some time within their history but remain quiescent for the vast majority of their history.

Turning this argument around, then, a fully self-consistent understanding of the AGN phenomenon requires a working mechanism to transform a hitherto-inactive galaxy into an active one. Such a mechanism would need to be able to force large amounts of material into the galaxy's central regions, as observations of bulges (particularly in massive galaxies) have consistently shown that the central bulges of galaxies represent an environment where many generations of star formation have exhausted the nuclear ISM, leaving it with much less material than would be required to support nuclear activity. An example of such an object would be our own galaxy, which possesses a supermassive black hole (Sag $A*$) that is, compared to the galaxies discussed in this chapter, remarkably inactive – estimates of its accretion luminosity are in the range of $\sim 10^{-8}$ times the Eddington rate.

The most consistent mechanism for such a transformation lies in the interaction and merging process that each galaxy undergoes throughout its history. Simulations of galactic interactions show that the merger process profoundly disturbs the matter distribution of each (e.g., Mihos and Hernquist 1996), eventually producing out of the merger a single aggregate elliptical galaxy (Barnes and Hernquist 1992) that takes several dynamical times (hundreds of millions of years) to fully relax. As the product galaxy is forming, material from the interstellar media of both galaxies falls toward the center of the aggregate. Moreover, at the same time, the central black holes of the two galaxies are also falling towards one another and will eventually merge. This latter point is actually important because it is difficult if not impossible to spin up a black hole in any way other than by merging it with another black hole that comes in with a different angular momentum and a considerable fraction of its mass.

It was for these reasons that Wilson and Colbert (1995) first proposed that a major merger of two disk galaxies could provide the mechanism for producing an active nucleus, particularly one that is radio-loud. The Wilson-Colbert hypothesis explains the strong correlation between radio-loud AGN and giant elliptical host galaxies (➜ Chap. 11) as well as the relative numbers of radio-quiet and radio-loud AGN, the latter being a by-product of the fact that most mergers are not of nearly equal mass objects, where the central black hole could be spun up. An unequal merger could then ignite activity, but not necessarily the ejection of jets, which appears to require a strong black hole spin. Under the merging scenario, an initial merger of disk galaxies at high redshift and their individual supermassive black holes would create a rapidly spinning product black hole and a luminous AGN. However, as the surrounding cluster or group of galaxies in which the AGN is embedded develops a hot ICM, a large fraction of the cluster galaxies will be stripped of most of their gas, thus ensuring that subsequent mergers are "dry," and provide less and less accretion power.

The merger scenario is supported by the discovery of twin, active black holes in the centers of two ultraluminous infrared galaxies (ULIRGs), NGC 6240 and Mrk 463 (Komossa et al. 2003; Bianchi et al. 2008). If the merger scenario is correct, these objects would represent an intermediate stage in the AGN development process, wherein the black holes are merging. Indeed, the host galaxies are clearly merger products, with their nuclei separated by a few kiloparsecs. A second line of evidence supporting the merger hypothesis is that HST imaging of the nearest compact symmetric objects (CSOs), which represent the very youngest of radio sources, appear to require that they underwent a major merger a few hundred million years ago (Perlman et al. 2001b) – long enough in the past that the merger of the black holes may either be in process or have taken place (e.g., Begelman et al. 1984) but also recently enough so that the nuclear medium would be expected to be dense enough to fuel an AGN. This delay of a few hundred million years is actually supported by the simulations, which show that the length of time required to force gas to the aggregate center is of the same order (di Matteo et al. 2005, 2008; Hopkins et al. 2005, 2006, 2008a, b; Springel et al. 2005a, b). It should be mentioned that in quasars, this phase might also correspond to the BAL stage, when the active nucleus is most dense. In such a case, one would expect that the luminosity of the AGN would peak when the object is very young and then decrease as the AGN ages and the amount of material available for accretion decreases. The latter mechanism has been used by Bicknell et al. (1997) to attempt to fit the luminosity function of CSOs as compared to FR II and FR I galaxies. Finally, some very recent work (Batcheldor et al. 2010) on M87 indicates that its supermassive black hole does not in fact lie in the galaxy's dynamical center but rather is offset from the center of the isophotes by 7 pc (projected), along the direction opposite to the jet. The most consistent explanation for this result is either that the black hole is displaced as a result of gravitational recoil resulting from the coalescence of binary supermassive black holes, or that it was accelerated by an intrinsically asymmetric jet. Since the latter hypothesis is not suggested by any other evidence, whereas the possibility of gravitational recoil is predicted by models of black hole mergers (e.g., Tichy and Marronetti 2007), which show that when two spinning black holes merge, gravitational waves can produce a kick of hundreds of km s^{-1} that damp out on Gyr timescales (Gualandris and Merritt 2008). Note, however, that due to the long damping timescale, any merger of spinning supermassive black holes can produce such an offset – not merely the one that might have been the activity's proximate cause.

As of the current writing, this merger scenario represents the leading hypothesis for the development of an active nucleus. Besides fleshing out various aspects of AGN properties and physics, it is therefore important to also test this mechanism for the development of nuclear activity.

References

Abdo, A. A., et al. 2009a, ApJS, 183, 46

Abdo, A. A., et al. 2009b, ApJ, 700, 597

Abdo, A. A., et al. 2009c, ApJ, 707, 55

Abraham, Z., Barres de Almeida, U., Dominici, T. P., & Caproni, A. 2007, MNRAS, 375, 171

Ackerman, et al. 2010, ApJ, 721, 1383

Aharonian, F., et al. 2006, Science, 314, 1424

Aharonian, F., et al. 2009, A&A, 502, 749

Alef, W., Gotz, M. M. A., Preuss, E., & Kellermann, K. I. 1988, A&A, 192, 53

Alef, W., Wu, S. Y., Preuss, E., Kellermann, K. I., & Qiu, Y. H. 1996, A&A, 308, 376

Allen, S. W., Dunn, R. J. H., Fabian, A. C., Taylor, G. B., & Reynolds, C. S. 2006, MNRAS, 372, 21

Aloy, M.-A., Marti, J.-M., Gómez, J.-L., Agudo, I., Müller, E., & Ibáñez, J.-M. 2003, ApJ, 589, L109

Antonucci, R. R. J. 1983, Nature, 303, 158

Antonucci, R. R. J. 1984, ApJ, 278, 499

Antonucci, R. R. J. 1993, ARAA, 31, 473

Antonucci, R. R. J., & Miller, J. S. 1985, ApJ, 297, 621

Antonucci, R. R. J., Hurt, T., & Kinney, A., 1994, ApJ, 430, 210

Antonucci, R. R. J., Geller, R., Goodrich, R. W., & Miller, J. S. 1996, ApJ, 472, 502

Arav, N., Barlow, T. A., Laor, A., & Blandford, R. D. 1997, MNRAS, 288, 1015

Arav, N., et al. 2007, ApJ, 658, 829

Arav, N., Barlow, T. A., Laor, A., Sargent, W. L. W., & Blandford, R. D. 1998, MNRAS, 297, 990

Arav, N., Kaastra, J., Kriss, G., Korista, K. T., Gabel, J., & Proga, D. 2005, ApJ, 620, 665

Aretxaga, I., & Terlevich, R. 1994, MNRAS, 269, 462

Armus, L., et al. 2006, ApJ, 640, 204

Arshakian, T. G., Léon-Tavares, J., Lobanov, A. P., Chavushyan, V. H., Shapovalova, A. I., Burenkov, A. N., & Zensus, J. A. 2010, MNRAS, 401, 1231

Babadzhanyants, M. K., & Hagen-Thorn, V. A. 1969, Astron. Tsirk., 526, 1

Balbus, S. A. 2003, ARAA, 41, 555

Barnes, T. G. 1968, ApL, 171, 1

Barnes, J. E., & Hernquist, L. 1992, ARAA, 30, 705

Batcheldor, D., Robinson, A., Axon, D. J., Perlman, E. S., & Merritt, D. 2010, ApJ (in press). arXiv: 1005.2173

Beckmann, V., Barthelmy, S. D., Courvoisier, T. J.-L., Gehrels, N., Soldi, S., Tueller, J., & Wendt, G. 2007, A&A, 475, 827

Begelman, M. C., & Sikora, M. 1987, ApJ, 322, 650

Begelman, M. C., Blandford, R. D., & Rees, M. J. 1984, Rev. Mod. Phys., 56, 255

Begelman, M. C., de Kool, M., & Sikora, M. 1991, ApJ, 382, 416

Beilicke, M., Benbow, W., Cornils, R., Heinzelmann, G., Horns, D., Raue, M., Ripken, J., & Tluzczykont, M. 2005, astro-ph/0504395

Benbow, W., Boisson, C., & Costamante, L. 2008, Proc. ICRC, 3, 1081, arXiv: 0709.4608

Bentz, M. C., Peterson, B. M., Pogge, R. W., Vestergaard, M., & Onken, C. A. 2006, ApJ, 644, 133

Bentz, M. C., et al. 2007, ApJ, 662, 205

Bentz, M. C., et al. 2009a, ApJ, 705, 199

Bentz, M. C., Peterson, B. M., Netzer, H., Pogge, R. W., & Vestergaard, M. 2009b, ApJ, 697, 160

Bentz, M. C., Peterson, B. M., Pogge, R., & Vestergaard, M. 2009c, ApJ, 694, L166

Bianchi, S., Chiaberge, M., Piconcelli, E., Guainazzi, M., & Matt, G. 2008, MNRAS, 386, 105

Bicknell, G. V., Dopita, M. A., & O'Dea, C. P. 1997, ApJ, 485, 112

Biretta, J. A., Zhou, F., & Owen, F. N. 1995, ApJ, 447, 582

Biretta, J. A., Sparks, W. B., & Macchetto, F. D. 1999, ApJ, 520, 621

Blandford, R. D., & Znajek, R. L. 1977, MNRAS, 179, 433

Bolton, J. G., Stanley, J. G., & Slee, O. B. 1949, Nature, 164, 101

Brenneman, L. W., & Reynolds, C. S. 2009, ApJ, 702, 1367

Brotherton, M. S., Tran, H. D., van Breugel, W., Dey, A., & Antonucci, R. 1997, ApJ, 487, L113

Brotherton, M. S., van Breugel, W., Smith, R. J., Boyle, B. J., Shanks, T., Croom, S. M., Miller, L., & Becker, R. H. 1998, ApJ, 505, L8

Brotherton, M. S., Laurent-Muehleisen, S. A., Becker, R. H., Gregg, M. D., Telis, G., White, R. L., & Shang, Z.-H. 2005, AJ, 130, 2006

Buchanan, C. L., Gallimore, J. F., O'Dea, C. P., Baum, S. A., Axon, D. J., Robinson, A., Elitzur, M., & Elvis, M. 2006, AJ, 132, 401

Cannon, R. D., Penston, M. V., & Penston, M. J. 1968, Nature, 217, 340

Cara, M., et al. 2010 (in prepare)

Chelouche, D., & Netzer, H. 2001, MNRAS, 326, 916

Cheung, C. C., Harris, D. E., & Stawarz, L. 2007, ApJ, 663, L65

Chini, R., Kruegel, E., Kreysa, E., & Gemuend, H.-P. 1989, A&A, 216, L5

Chitnis, V. R., Pendharkar, J. K., Bose, D., Agrawal, V. K., Rao, A. R., & Misra, R. 2009, ApJ, 698, 1207

Cirasuolo, M., Magliocchetti, M., Celotti, A., & Danese, L. 2003, MNRAS, 341, 993

Collin, S., Kawaguchi, T., Peterson, B. M., & Vestergaard, M. 2006, A&A, 456, 75

Collin, S., & Kawaguchi, T. 2004, A&A, 426, 797

Contopoulos, J. 1994, ApJ, 432, 508

Costantini, E., et al. 2007, A&A, 461, 121

Crenshaw, D. M., Kraemer, S. B., Boggess, A., Maran, S. P., Mushotzky, R. F., & Wu, C. 1999, ApJ, 516, 750

Crenshaw, D. M., Kraemer, S. B., & George, I. M. 2003, ARAA, 41, 117

Crenshaw, D. M., Kraemer, S. B., Gabel, J. R., Kaastra, J. S., & Steenbrugge, K. C. 2004, ApJ, 607, 794

Cristiani, S., Trentini, S., La Franca, F., Aretxaga, I., Andreani, P., Vio, R., & Gemmo, A. 1996, A&A, 306, 395

de Koff, S., Baum, S. A., Sparks, W. B., Biretta, J. A., Golombek, D., Macchetto, F. D., McCarthy, P., & Miley, G. K. 1996, ApJS, 110, 191

di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604

di Matteo, T., Colberg, J., Springel, V., Hernquist, L., & Sijaki, D. 2008, ApJ, 676, 33

di Serego Alighieri, S., Cimatti, A., & Fosbury, R. A. E. 1994, ApJ, 431, 123

Dondi, L., & Ghisellini, G. 1995, MNRAS, 273, 583

Efstathiou, A., & Rowan-Robinson, M. 1995, MNRAS, 273, 649

Ekers, R. D., Fanti, R., & Miley, G. K. 1983, A&A, 120, 297

Elitzur, M. 2007, in The Central Engine of Active Galactic Nuclei, eds. L. C. Ho & J.-M. Wang (San Francisco: ASP), 415. astro-ph/0612458

Elitzur, M., & Shlosman, I. 2006, ApJ, 648, L101

Elitzur, M., Nenkova, M., & Ivezic, Z. 2004, in The Neutral ISM in Starburst Galaxies, eds. S. Alto, S. Huttemeister, & A. Pedlar (San Francisco: ASP), 242. astro-ph/0309040

Fabian, A. C., Vaughan, S., Nandra, K. P., Iwasawa, K., Ballantyne, D. R., Lee, J. C., De Rosa, A., Turner, A., & Young, A. J. 2002, MNRAS, 335, L1

Fanaroff, B. L., & Riley, J. M. 1974, MNRAS, 167, 31

Fath, E. A. 1909, Lick Obs. Bull. 5, 71

Ferguson, J. W., Korista, K. T., Baldwin, J. A., & Ferland, G. J. 1997, ApJ, 487, 122

Ferland, G. J., Korista, K. T., Verner, D. A., Ferguson, J. W., Kingdon, J. B., & Verner, E. M. 1998, PASP, 110, 761

Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9

Filippenko, A. V. 1989, AJ, 97, 726

Fossati, G., et al. 2008, ApJ, 677, 906

Francis, P. J. 1996, PASA, 13, 212

Gabel, J. R., et al. 2005, ApJ, 631, 741

Gallagher, S. C., Brandt, W. N., Chartas, G., Priddey, R., Garmire, G. P., & Sambruna, R. M. 2006, ApJ, 644, 709

Georganopoulos, M., Kazanas, D., Perlman, E. S., & Stecker, F. W. 2005, ApJ, 625, 656

Georganopoulos, M., Perlman, E. S., Kazanas, D., & Wingert, B. 2006, in Blazar Variability Workshop II: Entering the GLAST Era, ed. H. R. Miller, K. Marshall, J. R. Webb, & M. F. Aller (San Francisco: ASP), 178

George, I. M., Turner, T. J., Mushotzky, R. F., Nandra, K., & Netzer, H. 1998, ApJ, 503, 174

Ghisellini, G., Tavecchio, F., & Chiaberge, M. 2005, A&A, 432, 401

Giovannini, G., Taylor, G. B., Arbizzani, E., Bondi, M., Cotton, W. D., Feretti, L., Lara, L., & Venturi, T. 1999, ApJ, 522, 101

Giovannini, G., Giroletti, M., & Taylor, G. B. 2007, A&A, 474, 409

Giustini, M., Cappi, M., & Vignali, C. 2008, A&A, 491, 425

Gliozzi, M., Papadakis, I. E., Eracleous, M., Sambruna, R. M., Ballantyne, D. R., Braito, V., & Reeves, J. N. 2009, ApJ, 703, 1021
Gomez, J.-L., Marscher, A. P., Alberdi, A., Marti, J. M. A., & Ibanez, J. M. A. 1998, ApJ, 499, 221
Goodrich, R. W., & Cohen, M. H. 1992, ApJ, 391, 623
Granato, G. L., & Danese, L. 1994, MNRAS, 268, 235
Green, P. J., et al., 2009, ApJ, 690, 644
Green, P. J., & Mathur, S. 1996, ApJ, 462, 637
Greene, J. E., & Ho, L. C. 2007, ApJ, 667, 131
Greenhill, L. J., Gwinn, C. R., Antonucci, R. R. J., & Barvainis, R. 1996, ApJ, 472, L21
Greenstein, J. L., & Matthews, T. 1963, Nature, 197, 1041
Greenstein, J. L., & Schmidt, M. 1963, Nature, 197, 1041
Greenstein, J. L., & Schmidt, M. 1963, ApJ, 1
Guainazzi, M., Matt, G., Brandt, W. N., Antonelli, L. A., Barr, P., & Bassani, L. 2000, A&A, 356, 463
Gualandris, A., & Merritt, D. 2008, ApJ, 678, 780
Hardcastle, M. J., Birkinshaw, M., & Worrall, D. M. 2001, MNRAS, 326, 1499
Hardcastle, M. J., Harris, D. E., Worrall, D. M., & Birkinshaw, M. 2004, ApJ, 612, 729
Hardee, P. E. 2003, ApJ, 597, 798
Hardee, P. E., & Rosen, A. 2002, ApJ, 576, 204
Hardee, P. E., Hughes, P. A., Rosen, A., & Gomez, E. A. 2001, ApJ, 555, 744
Hardee, P. E., Walker, R. C., & Gómez, J. L. 2005, ApJ, 620, 646
Hardee, P. E., Mizuno, Y., & Nishikawa, K. 2007, Ap&SS, 311, 281
Harris, D. E., & Krawczynski, H. 2002, ApJ, 566, 244
Harris, D. E., & Krawczynski, H. 2006, ARAA, 44, 463
Harris, D. E., Biretta, J. A., Junor, W., Perlman, E. S., Sparks, W. B., & Wilson, A. S. 2003, ApJ, 586, L41
Harris, D. E., Cheung, C. C., Stawarz, L., Biretta, J. A., & Perlman, E. S. 2009, ApJ, 699, 305
Hazard, C., Mackey, M. B., & Shimmins, A. J. 1963, Nature, 197, 1037
Hicks, E. K. S., & Malkan, M. A. 2008, ApJS, 174, 31
Hopkins, P. F., Hernquist, L., Cox, T. J., Di Matteo, T., Martini, P., Robertson, B., & Springel, V. 2005, ApJ, 630, 705
Hopkins, P. F., Hernquist, L., Cox, T. J., Di Matteo, T., Robertson, B., & Springel, V. 2006, ApJS, 163, 1
Hopkins, P. F., Cox, T. J., Keres, D., & Hernquist, L. 2008a, ApJS, 175, 356
Hopkins, P. F., Hernquist, L., Cox, T. J., & Keres, D. 2008b, ApJS, 175, 390
Hubble, E. 1926, ApJ, 64, 321
Hughes, P. A., Duncan, C., & Mioduszewski, A. 1996, in Energy Transport in Radio Galaxies and Quasars, eds. P. E. Hardee, A. H. Bridle, & J. A. Zensus (San Francisco: ASP), 137
Hughes, P. A., Miller, M. A., & Duncan, G. C. 2002, ApJ, 572, 713
Impey, C. D., & Neugebauer, G. 1988, AJ, 95, 307
Ivezic, Z., et al. 2002, AJ, 124, 2364
Ivezic, Z., et al. 2004, in AGN Physics with the Sloan Digital Sky Survey, eds. G. T. Richards & P. B. Hall, ASP Conference Series, Vol. 311, 347. astro-ph/0310569
Iwasawa, K., & Taniguchi, Y. 1993, ApJ, 413, L15
Jackson, N., & Tadhunter, C. N. 1993, A&A, 272, 105
Jaffe, W., et al. 2004, Nature, 429, 407
Jansky, K. G. 1932, Proc. IRE, 20, 1920
Jansky, K. G. 1933, Proc. IRE, 21, 1387
Jansky, K. G. 1935, Proc. IRE, 23, 1158
Jiang, L., Fan, X., Ivezic, Z., Richards, G., Schneider, D. P., Strauss, M. A., & Kelley, B. C. 2007, ApJ, 656, 680
Jorstad, S. G., Marscher, A. P., Mattox, J. R., Wehrle, A. E., Bloom, S. D., & Yurchenko, A. V. 2001, ApJS, 134, 181
Jorstad, S. G., et al. 2005, AJ, 130, 1418
Joshi, et al. 2011, arXiv: 1111.0984
Jourdain, E., et al. 1993, ApJ, 412, 586
Kallman, T., & Bautista, M. 2001, ApJ Suppl., 133, 221
Kaspi, S., et al. 2002, ApJ, 574, 643
Kaspi, S., Brandt, W. N., Maoz, D., Netzer, H., Schneider, D. P., & Shemmer, O. 2007, ApJ, 659, 997
Kataoka, J., et al. 2007, PASJ, 59, 279
Kellermann, K. I., Sramek, R., Schmidt, M., Shaffer, D. B., & Green, R. 1989, AJ, 98, 1195
Kelly, B. C., & Bechtold, J. 2007, ApJS, 168, 1
Kinman, T. D. 1968, Science, 162, 1081
Kniffen, D. A., et al. 1993, ApJ, 411, 133
Komossa, S., Burwitz, V., Hasinger, G., Predehl, P., Kaastra, J. S., & Ikebe, Y. 2003, ApJ, 582, L15
Koratkar, A., & Blaes, O. 1999, PASP, 111, 1
Körding, E. G., Jester, S., & Fender, R. 2006, MNRAS, 372, 1366
Körding, E. G., Jester, S., & Fender, R. 2008, MNRAS, 383, 277
Korista, K. T., & Goad, M. R. 2000, ApJ, 536, 284
Korista, K. T., & Goad, M. R. 2004, ApJ, 606, 749
Korista, K. T., et al. 1995, ApJS, 97, 285
Kraemer, S. B., Crenshaw, D. M., & Gabel, J. R. 2001a, ApJ, 557, 30
Kraemer, S. B., et al. 2001b, ApJ, 551, 67
Krichbaum, T. P., Graham, D. A., Bremer, M., Alef, W., Witzel, A., Zensus, J. A., & Eckart, A., 2006, in The Universe Under the Microscope – Astrophysics at High Angular Resolution, Journal of Physics: Conference Series, Vol. 54, 328

Krolik, J. H., 1998, Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment, (Princeton: Princeton University Press)

Larsson, J., Levan, A. J., Davies, M. B., & Fruchter, A. S. 2007, MNRAS, 376, 1285

Leahy, J. P., & Perley, R. A. 1991, AJ, 102, 537

Li, Z.-Y., Chiueh, T., & Begelman, M. C. 1992, ApJ, 394, 459

Longair, M. S. 1994, High Energy Astrophysics, vol. 2: Stars and the Interstellar Medium (Cambridge: Cambridge University Press)

Lynden-Bell, D. 1969, Nature, 223, 690

Magorrian, J., Tremaine, S., Richstone, D., Bender, R., Bower, G., Dressler, A., Faber, S. M., Gebhardt, K., Green, R., Grilllmair, C., Kormendy, J., & Lauer, T. 1998, AJ, 115, 2285

Maiolino, R., & Risaliti, G. 2007, in The Central Engine of Active Galactic Nuclei, ed. L. C. Ho, & J.-M. Wang (San Francisco: ASP), 447

Maraschi, L., Ghisellini, G., & Celotti, A. 1992, ApJ, 397, L5

Maraschi, L., et al. 1994a, ApJ, 435, L91

Marshall, H. L., Miller, B. P., Davis, D. S., Perlman, E. S., Wise, M., Canizares, C. R., & Harris, D. E. 2002, ApJ, 564, 683

Marshall, K., Ryle, W. T., Miller, H. R., Marscher, A. P., Jorstad, S. G., Chicka, B., & McHardy, I. M. 2009, ApJ, 696, 601

Mason, R. E., Geballe, T. R., Packham, C., Levenson, N. A., Elitzur, M., Fisher, R. S., & Perlman, E. 2006, ApJ, 640, 612

Matt, G., et al. 1997, A&A, 327, L13

McGill, K. L., Woo, J.-H., Treu, T., & Malkan, M. 2008, ApJ, 673, 703

McNamara, A. L., Kuncic, Z., & Wu, K. 2009, MNRAS, 395, 1507

Mehta, K. T., Georganopoulos, M., Perlman, E. S., Padgett, C. A., & Chartas, G. 2009, ApJ, 690, 1706

Meier, D. L., Koide, S., & Uchida, Y. 2001, Science, 291, 84

Meisenheimer, K., et al. 2007, A&A, 471, 453

Melia, F. 2009, High-Energy astrophysics, (Princeton)

Mihos, J. C., & Hernquist, L. 1996, ApJ, 464, 641

Miller, J. S., Goodrich, R. W., & Mathews, W. G. 1991, ApJ, 378, 47

Moore, J. H. 1915, PASP, 21, 192

Mutel, R., Phillips, R., Su, B., & Buciferro, R. 1990, ApJ, 352, 81

Nandra, K., George, I. M., Mushotzky, R. F., Turner, T. J., & Yaqoob, T. 1997, ApJ, 488, L91

Nenkova, M., Sirocky, M. M., Ivezic, Z., & Elitzur, M. 2008a, ApJ, 685, 147

Nenkova, M., Sirocky, M. M., Nikutta, R., Ivezic, Z., & Elitzur, M. 2008b, ApJ, 685, 160

Netzer, H. 1990, in Active Galactic Nuclei, ed. T. Courvoisier, & M. Mayor (Berlin: Springer)

O'Neill, S. M., Tregillis, I. L., Jones, T. W., & Ryu, D. 2005, ApJ, 633, 717

Ogle, P. M., Cohen, M. H., Miller, J. S., Tran, H. D., Goodrich, R. W., & Martel, A. R. 1999, ApJ Suppl., 125, 1

Onken, C. A., & Kollmeier, J. A. 2008, ApJ, 689, 13

Onken, C. A., et al. 2007, ApJ, 670, 105

Oke, J. B. 1963, Nature, 197, 1040

Osterbrock, D. E. 1999, ApJ, 525, 337

Osterbrock, D. E., & Ferland, G. J. 2006, in Astrophysics of Gaseous Nebulae and Active Galactic Nuclei (2nd ed.; Sausalito: University Science Books)

Pacholczyk, A. G., & Wyemann, R. T. 1968, AJ, 73, 850

Packham, C., Young, S., Hough, J. H., Axon, D. J., & Bailey, J. A. 1997, MNRAS, 288, 375

Packham, C., Radomski, J. T., Roche, P. F., Aitken, D. K., Perlman, E. S., Alonso-Herrero, A., Colina, L., & Telesco, C. M. 2005, ApJ, 618, L17

Padovani, P., Giommi, P., Landt, H., & Perlman, E. S. 2007, ApJ, 662, 182

Pease, F. G. 1915, PASP, 27, 133

Perlman, E. S., et al. 2011, ApJ, 743, 119

Perlman, E. S., & Wilson, A. S. 2005, ApJ, 627, 140

Perlman, E. S., Biretta, J. A., Zhou, F., Sparks, W. B., & Macchetto, F. D. 1999, AJ, 117, 2185

Perlman, E. S., Biretta, J. A., Sparks, W. B., Macchetto, F. D., & Leahy, J. P. 2001a, ApJ, 551, 206

Perlman, E. S., Stocke, J. T., Conway, J. E., & Reynolds, C. 2001b, AJ, 122, 536

Perlman, E. S., Addison, B., Georganopoulos, M., Wingert, B., & Graff, P. 2008, in Blazar Variability Across the Electromagnetic Spectrum, http://pos.sissa.it,p.9,arXiv:0807.2119

Peterson, B. M. 1993, PASP, 105, 247

Peterson, B. M., et al. 2005, ApJ, 632, 799; Erratum: 2006, ApJ, 741, 638

Pier, E. A., & Krolik, J. H. 1992, ApJ, 399, L23

Poncelet, A., & Perrin, G., Sol, H. 2006, A&A, 450, 483

Poncelet, A., Doucet, C., Perrin, G., Sol, H., & Lagage, P. O. 2007, A&A, 472, 823

Poncelet, A., Sol, H., & Perrin, G. 2008, A&A, 481, 305

Proga, D., Stone, J. M., & Kallman, T. R. 2000, ApJ, 543, 686

Radomski, J. T., Piña, R. K., Packham, C., Telesco, C. M., De Buizer, J. M., Fisher, R. S., & Robinson, A. 2003, ApJ, 587, 117

Radomski, J. T., et al. 2008, ApJ, 681, 141

Reber, G. 1940a, Proc. IRE, 28, 68

Reber, G. 1940b, ApJ, 100, 279

Rees, M. J. 1966, Nature, 211, 468

Reeves, J. N., et al. 2007, PASJ, 59, 301

Reichard, T. A., Richards, G. T., Hall, P. B., Schneider, D. P., Vanden Berk, D. E., Fan, X.-H., York, D. G., Knapp, G. R., & Brinkmann, J. 2003a, AJ, 126, 2594

Reichard, T. A., Richards, G. T., Schneider, D. P., Hall, P. B., Tolea, A., Krolik, J. H., Tsvetanov, Z., Vanden Berk, D. E., York, D. G., Knapp, G. R., Gunn, J. E., & Brinkmann, J. 2003b, AJ, 125, 1711

Reynolds, C. S. 1996, PhD Thesis, Cambridge University

Reynolds, C. S., & Fabian, A. C. 2008, ApJ, 675, 1048

Reynolds, C. S., Fabian, A. C., Brenneman, L. W., Miniutti, G., Uttley, P., & Gallo, L. C. 2009, MNRAS, 397, 21

Reynolds, C. S., & Miller, M. C. 2009, ApJ, 692, 869

Reynolds, C. S., & Nowak, M. A. 2003, Phys. Rep., 389, 466

Risaliti, G. 2002, in Inflows, Outflows & Reprocessing Around Black Holes, Proceedings of the 5th Italian AGN Meeting, http://www.unico.it/ilaria/AGNS/proceedings.html,p.111

Risaliti, G., & Elvis, M. 2004, in Supermassive Black Holes in the Distant Universe, ed. A. J. Barger (Dordrecht: Kluwer), 187, astro-ph/0403618

Risaliti, G., & Elvis, M. 2010, A&A, 516, 89

Roche, P. F., Packham, C., Telesco, C. M., Radomski, J. T., Alonso-Herrero, A., Aitken, D. K., Colina, L., & Perlman, E. S. 2006, MNRAS, 267, 1689

Roche, P. F., Packham, C., Aitken, D. K., & Mason, R. E. 2007, MNRAS, 375, 99

Ross, H. N. 1970, Nature, 226, 431

Rybicki, G. H., & Lightman, A. P. 1986, Radiative Processes in Astrophysics (New York: Wiley)

Ryle, M. & Smith, F. G. 1948, Nature, 162, 462

Schmidt, M. 1963, Nature, 197, 1040

Scott, J. E., et al. 2009, ApJ, 694, 438

Scott, J. E., Kriss, G. A., Brotherton, M., Green, R. F., Hutchings, J., Shull, J. M., & Zheng, W. 2004, ApJ, 615, 135

Selove, D. M. 1969, ApJ, 158, 19

Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337

Shang, Z., et al. 2005, ApJ, 619, 41

Shemmer, O., Brandt, W. N., Gallagher, S. C., Vignali, C., Boller, T., Chartas, G., & Comastri, A. 2005, AJ, 130, 2522

Shields, G. 1999, PASP, 111, 661

Seyfert, C. K. 1943, ApJ, 98, 20

Sikora, M., Madejski, G., Moderski, R., & Poutanen, J. 1997, ApJ, 484, 108

Sikora, M., Stawarz, L., & Lasota, J.-P. 2007, ApJ, 658, 815

Slipher, V. M. 1917, Lowell Obs. Bull., 3, 59

Springel, V., di Matteo, T., & Hernquist, L. 2005a, MNRAS, 361, 776

Springel, V., et al. 2005b, Nature, 435, 629

Stocke, J. T., Morris, S. L., Weymann, R. J., & Foltz, C. B. 1992, ApJ, 396, 487

Stone, J. M., & Balbus, S. A. 1996, ApJ, 464, 364

Stull, M. A. 1970, Nature, 225, 832

Suganuma, M., Kobayashi, Y., Yoshii, Y., Minezaki, T., Tomita, H., Aoki, T., Enya, K., Koshida, S., & Peterson, B. A. 2007, in The Central Engine of Active Galactic Nuclei, ed. L. C. Ho, & J.-M. Wang (San Francisco: ASP), 462

Tang, S. M., Zhang, S. N., & Hopkins, P. F. 2007, MNRAS, 377, 1133

Tchekhovskoy, A., Narayan, R., & McKinney, J. C. 2010, ApJ, 711, 50

Tchekhovskoy, A., Narayan, R., & McKinney, J. C. 2011, ApJ, 711, 50

Terlevich, R., Tenorio-Tagle, G., Franco, J., & Melnick, J. 1992, MNRAS, 255, 713

Tichy, W., & Marronetti, P. 2007, Phys. Rev. D., 76, 061502

Tristram, K. R. W., et al. 2007, A&A, 474, 837

Tristram, K. R. W., et al. 2009, A&A, 502, 67

Tsvetanov, Z. I., Hartig, G. F., Ford, H. C., Dopita, M. A., Kriss, G. A., Pei, Y. C., Dressel, L. L., & Harms, R. J. 1998, ApJ, 493, L83

Turnshek, D. A. 1987, in QSO Absorption Lines, STScI Symposium Series 2, ed. J. C. Blades, D. A. Turnshek, & C. A. Norman (Cambridge: Cambridge University Press), 17

Uchiyama, Y., & Coppi, P. 2010 (in prepare)

Ulrich, M.-H., Maraschi, L., & Urry, C. M. 1997, ARAA, 35, 445

Urry, C. M., & Padovani, P. 1995, PASP, 107, 803

Valtaoja, E., Terasranta, H., Urpo, S., Nesterov, N. S., Lainela, M., & Valtonen, M. 1992, A&A, 254, 80

Van den Berk, D. E., et al. 2001, AJ, 122, 549

Veilleux, S., & Osterbrock, D. 1987, ApJS, 63, 295

Veron-Cetty, M. P., Veron, P. 2000, A&AR, 10, 81

Vestergaard, M., & Osmer, P. S. 2009, ApJ, 699, 800

Vestergaard, M., & Peterson, B. M. 2006, ApJ, 641, 689

Vestergaard, M., Fan, X., Tremonti, C. A., Osmer, P. S., & Richards, G. T. 2008, ApJ, 674, L1

Vlahakis, N., & Königl, A. 2003, ApJ, 596, 1104

Vlahakis, N., & Königl, A. 2004, ApJ, 605, 656

Walker, R. C., Ly, C., Junor, W., & Hardee, P. 2009, in Approaching Micro-Arcsecond Resolution with VSOP-2: Astrophysics and Techniques, ed. Y. Hagiwara, E. Fomalont, M. Tsuboi, & Y. Murata (Astronomical Society of the Pacific: San Francisco), 227

Wang, H.-Y., Wang, T.-G., & Wang, J.-X. 2005, ApJ, 634, 149

Wang, J.-G., Dong, X.-B., Wang, T.-G., Ho, L. C., Yuan, W., Wang, H., Zhang, K., Zhang, S., & Zhou, H. 2009, ApJ, 707, 1334

Wehrle, A. E., et al. 1998, ApJ, 497, 178

Weymann, R. J., Morris, S. L., Foltz, C. B., & Hewett, P. C. 1991, ApJ, 373, 23

Wilson, A. S., & Colbert, E. J. M. 1995, ApJ, 438, 62

Wilson, A. S., Braatz, J. A., Heckman, T. M., Krolik, J. H., & Miley, G. K. 1993, ApJ, 419, L61

Wu, Y., Charmandaris, V., Huang, J., Spinoglio, L., & Tommasin, S. 2009, ApJ, 701, 658

Yip, C. W., et al. 2004, AJ, 128, 2603

Young, S., Packham, C., Mason, R. E., Radomski, J. T., & Telesco, C. M. 2007, MNRAS, 378, 888

Zaitseva, G. V., & Lyutyi, V. M. 1969, Astron. Zh., 526, 1

Zhang, S.-Y., Bian, W.-H., & Huang, K.-L. 2008, A&A, 488, 113

Zheng, W., Kriss, G. A., Telfer, R. C., Grimes, J. P., & Davidsen, A. F. 1998, ApJ, 492, 855

8 The Large-Scale Structure of the Universe

Alison L. Coil
Department of Physics, University of California, San Diego, CA, USA

1 Historical Background *388*

2 The Two-Point Correlation Function *391*

3 Angular Clustering *393*

4 Real and Redshift Space Clustering *395*

5 Galaxy Bias *398*

6 The Dependence of Clustering on Galaxy Properties *400*
6.1 Luminosity Dependence 400
6.2 Color and Spectral-Type Dependence 401
6.3 Redshift Space Distortions 403

7 The Evolution of Galaxy Clustering *405*

8 Halo Model Interpretation of $\xi(r)$ *406*
8.1 Estimating the Mean Halo Mass from the Bias 406
8.2 Halo Occupation Distribution Modeling 407
8.3 Interpreting the Luminosity and Color Dependence of Galaxy Clustering 410
8.4 Interpreting the Evolution of Galaxy Clustering 411

9 Voids and Filaments *413*
9.1 Higher-Order Clustering Measurements 413
9.2 Voids 414
9.2.1 Void and Void Galaxy Properties 415
9.2.2 Void Probability Function 417
9.3 Filaments 418

10 Summary and Future *419*

Acknowledgments *420*

References *420*

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_8, © Springer Science+Business Media Dordrecht 2013

Abstract: Galaxies are not uniformly distributed in space. On large scales, the Universe displays coherent structure, with galaxies residing in groups and clusters on scales of ~1–3 h^{-1} Mpc, which lie at the intersections of long filaments of galaxies that are >10 h^{-1} Mpc in length. Vast regions of relatively empty space, known as voids, contain very few galaxies and span the volume in between these structures. This observed large-scale structure depends both on cosmological parameters and on the formation and evolution of galaxies. Using the two-point correlation function, one can trace the dependence of large-scale structure on galaxy properties, such as luminosity, color, stellar mass, and track its evolution with redshift. Comparison of the observed galaxy clustering signatures with dark matter simulations allows one to model and understand the clustering of galaxies and their formation and evolution within their parent dark matter halos. Clustering measurements can determine the parent dark matter halo mass of a given galaxy population, connect observed galaxy populations at different epochs, and constrain cosmological parameters and galaxy evolution models. This chapter describes the methods used to measure the two-point correlation function in both redshift and real space, presents the current results of how the clustering amplitude depends on various galaxy properties, and discusses quantitative measurements of the structures of voids and filaments. The interpretation of these results with current theoretical models is also presented.

1 Historical Background

Large-scale structure is defined as the structure or inhomogeneity of the Universe on scales larger than that of a galaxy. The idea of whether galaxies are distributed uniformly in space can be traced to Edwin Hubble, who used his catalog of 400 "extragalactic nebulae" to test the homogeneity of the Universe (Hubble 1926), finding it to be generally uniform on large scales. In 1932, the larger Shapley-Ames catalog of bright galaxies was published (Shapley and Ames 1932), in which the authors note "the general unevenness in distribution" of the galaxies projected onto the plane of the sky and the roughly factor of 2 difference in the numbers of galaxies in the northern and southern galactic hemispheres. Using this larger statistical sample, Hubble (1934) noted that on angular scales less than ~10°, there is an excess in the number counts of galaxies above what would be expected for a random Poisson distribution, though the sample follows a Gaussian distribution on larger scales. Hence, while the Universe appears to be homogeneous on the largest scales, on smaller scales it is clearly clumpy.

Measurements of large-scale structure took a major leap forward with the Lick galaxy catalog produced by Shane and Wirtanen (1967), which contained information on roughly a million galaxies obtained using photographic plates at the 0.5-m refractor at Lick Observatory. Seldner et al. (1977) published maps of the counts of galaxies in angular cells across the sky (see ➋ *Fig. 8-1*), which showed in much greater detail that the projected distribution of galaxies on the plane of the sky is not uniform. The maps display a rich structure with a foamlike pattern containing possible walls or filaments with long strands of galaxies, clusters, and large empty regions (see ➋ Chap. 6). The statistical spatial distribution of galaxies from this catalog and that of Zwicky et al. (1968) was analyzed by Jim Peebles and collaborators in a series of papers (e.g., Peebles 1975) that showed that the angular two-point correlation function (defined below) roughly follows a power-law distribution over angular scales of ~0.1–5°. In these papers it was discovered that the clustering amplitude is lower for fainter galaxy populations, which

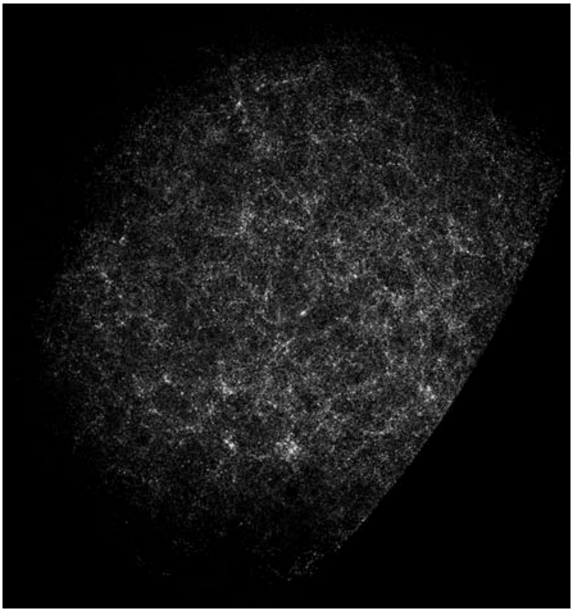


Fig. 8-1

Angular distribution of counts of galaxies brighter than *B* ~ 19 on the plane of the sky reconstructed from the Lick galaxy catalog (from Seldner et al. 1977, reproduction created by Ed Groth). This image shows the number of galaxies observed in 10′ × 10′ cells across the northern galactic hemisphere, where brighter cells contain more galaxies. The northern galactic pole is at the center, with the galactic equator at the edge. The distribution of galaxies is clearly not uniform; clumps of galaxies are seen in *white*, with very few galaxies observed in the dark regions between

likely arises from larger projection effects along the line of sight. As faint galaxies typically lie at larger distances, the projected clustering integrates over a wider volume of space and therefore dilutes the effect.

These results in part spurred the first large scale redshift surveys, which obtained optical spectra of individual galaxies in order to measure the redshifts and spatial distributions of large galaxy samples. Pioneering work by Gregory and Thompson (1978) mapped the three-dimensional spatial distribution of 238 galaxies around and towards the Coma/Abell 1367 supercluster. In addition to surveying the galaxies in the supercluster, they found that in the foreground at lower redshift there were large regions (>20 h^{-1} Mpc, shown well in their Figure 2a) with no galaxies, which they termed "voids." Joeveer et al. (1978) used redshift information from Sandge and Tammann (1975) to map the three-dimensional distribution of galaxies on large scales in the southern galactic hemisphere. They mapped four separate volumes that included clusters as well as field galaxies and showed that across large volumes, galaxies are clearly clustered in three dimensions and often form "chains" of clusters (now recognized as filaments).

Two additional redshifts surveys were the KOS survey (Kirshner et al. 1978) and the original CfA survey (Davis et al. 1982). The KOS survey measured redshifts for 164 galaxies brighter than magnitude 15 in eight separate fields on the sky, covering a total of 15 deg^2. Part of the motivation for the survey was to study the three dimensional spatial distribution of galaxies, about which the authors note that "although not entirely unexpected, it is striking how strongly clustered our galaxies are in velocity space," as seen in strongly peaked one dimensional redshift histograms in each field.

The original CfA survey, completed in 1982, contained redshifts for 2,400 galaxies brighter than magnitude 14.5 across the north and south galactic poles, covering a total of 2.7 steradians. The major aims of the survey were cosmological and included quantifying the clustering of galaxies in three-dimensions. This survey produced large area, moderately deep three dimensional maps of large-scale structure (see ➋ *Fig. 8-2*), in which one could identify galaxy clusters, voids, and an apparent "filamentary connected structure" between groups of galaxies, which the authors caution could be random projections of distinct structures (Davis et al. 1982). This paper also performed a comparison of the so-called complex topology of the large-scale structure seen in the galaxy distribution with that seen in N-body dark matter simulations, paving the way for future studies of theoretical models of structure formation.

The second CfA redshift survey, which ran from 1985 to 1995, contained spectra for ~5,800 galaxies and revealed the existence of the so-called Great Wall, a supercluster of galaxies that extends over 170 h^{-1} Mpc, the width of the survey (Geller and Huchra 1989). Large underdense voids were also commonly found, with a density 20% of the mean density.

Redshift surveys have rapidly progressed with the development of multi-object spectrographs, which allow simultaneous observations of hundreds of galaxies, and larger telescopes,

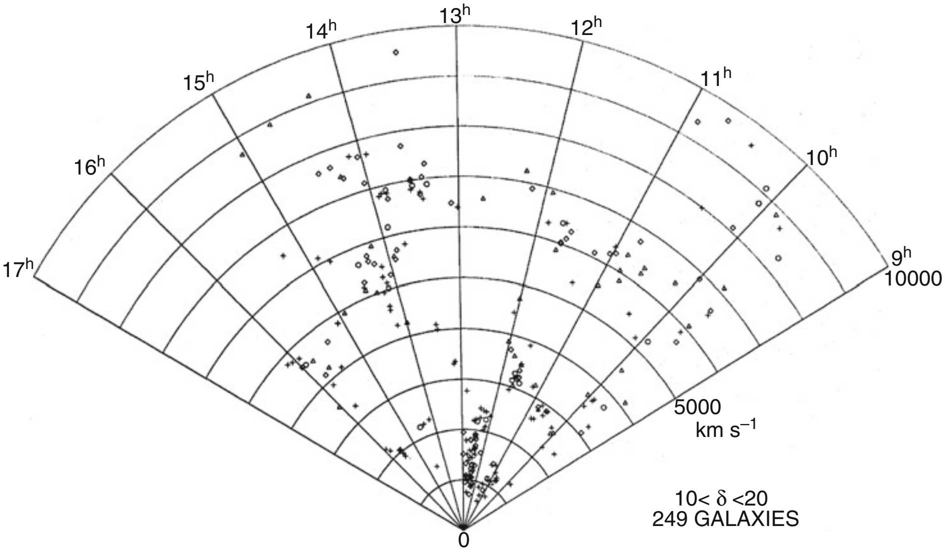


Fig. 8-2

Distribution of galaxies in redshift space from the original CfA galaxy redshift survey (from Davis et al. 1982). Plotted are 249 galaxies as a function of observed velocity (corresponding to a given redshift) versus right ascension for a wedge in declination of 10°< δ < 20°

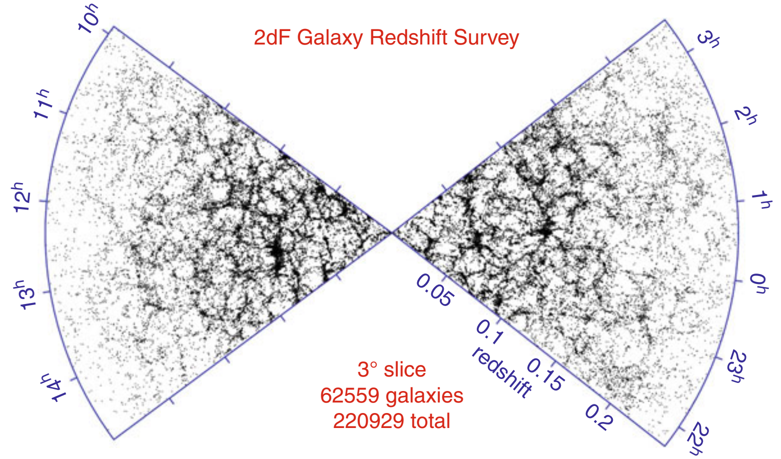


Fig. 8-3
The spatial distribution of galaxies as a function of redshift and right ascension (projected through 3° in declination) from the 2dF Galaxy Redshift Survey (from Colless 2004)

which allow deeper surveys of both lower luminosity nearby galaxies and more distant, luminous galaxies. At present, the largest redshift surveys of galaxies at low redshift are the Two Degree Field Galaxy Redshift Survey (2dFGRS, Colless et al. 2001) and the Sloan Digital Sky Survey 9SDSS, York et al. 2000), which cover volumes of ~4 × 10^7 h^3 Mpc^{-3} and ~2 × 10^8 h^3 Mpc^{-3} with spectroscopic redshifts for ~220,000 and a million galaxies, respectively. These surveys provide the best current maps of large-scale structure in the Universe today (see *Fig. 8-3*), revealing a spongelike pattern to the distribution of galaxies (Gott et al. 1986). Voids of ~10 h^{-1} Mpc are clearly seen, containing very few galaxies. Filaments stretching greater than 10 h^{-1} Mpc surround the voids and intersect at the locations of galaxy groups and clusters.

The prevailing theoretical paradigm regarding the existence of large-scale structure is that the initial fluctuations in the energy density of the early Universe, seen as temperature deviations in the cosmic microwave background, grow through gravitational instability into the structure seen today in the galaxy density field. The details of large-scale structure – the sizes, densities, and distribution of the observed structure – depend both on cosmological parameters such as the matter density and dark energy, as well as on the physics of galaxy formation and evolution. Measurements of large-scale structure can therefore constrain both cosmology (see Chap. 13) and galaxy evolution physics.

2 The Two-Point Correlation Function

In order to quantify the clustering of galaxies, one must survey not only galaxies in clusters (see Chap. 6) but rather the entire galaxy density distribution, from voids to superclusters. The most commonly used quantitative measure of large-scale structure is the galaxy two-point correlation function, $\xi(r)$, which traces the amplitude of galaxy clustering as a function of scale.

$\xi(r)$ is defined as a measure of the excess probability dP, above what is expected for an unclustered random Poisson distribution, of finding a galaxy in a volume element dV at a separation r from another galaxy,

$$dP = n[1 + \xi(r)]dV, \tag{8.1}$$

where n is the mean number density of the galaxy sample in question (Peebles 1980). Measurements of $\xi(r)$ are generally performed in comoving space, with r having units of h^{-1} Mpc. The Fourier transform of the two-point correlation function is the power spectrum, which is often used to describe density fluctuations observed in the cosmic microwave background.

To measure $\xi(r)$, one counts pairs of galaxies as a function of separation and divides by what is expected for an unclustered distribution. To do this, one must construct a "random catalog" that has the identical three-dimensional coverage as the data – including the same sky coverage and smoothed redshift distribution – but is populated with random distribution points. The ratio of pairs of galaxies observed in the data relative to pairs of points in the random catalog is then used to estimate $\xi(r)$. Several different estimators for $\xi(r)$ have been proposed and tested. An early estimator that was widely used is from Davis and Peebles (1983),

$$\xi = \frac{n_R}{n_D}\frac{\mathrm{DD}}{\mathrm{DR}} - 1, \tag{8.2}$$

where DD and DR are counts of pairs of galaxies (in bins of separation) in the data catalog and between the data and random catalogs, and n_D and n_R are the mean number densities of galaxies in the data and random catalogs. Hamilton (1993) later introduced an estimator with smaller statistical errors,

$$\xi = \frac{\mathrm{DD\ RR}}{(\mathrm{DR})^2} - 1, \tag{8.3}$$

where RR is the count of pairs of galaxies as a function of separation in the random catalog. The most commonly used estimator is from Landy and Szalay (1993),

$$\xi = \frac{1}{\mathrm{RR}}\left[\mathrm{DD}\left(\frac{n_R}{n_D}\right)^2 - 2\mathrm{DR}\left(\frac{n_R}{n_D}\right) + \mathrm{RR}\right]. \tag{8.4}$$

This estimator has been shown to perform as well as the Hamilton estimator (➋ 8.3), and while it requires more computational time, it is less sensitive to the size of the random catalog and handles edge corrections well, which can affect clustering measurements on large scales (Kerscher et al. 2000).

As can be seen from the form of the estimators given above, measuring $\xi(r)$ depends sensitively on having a random catalog which accurately reflects the various spatial and redshift selection affects in the data. These can include effects such as edges of slitmasks or fiber plates, overlapping slitmasks or plates, gaps between chips on the CCD, and changes in spatial sensitivity within the detector (i.e., the effective radial dependence within X-ray detectors). If one is measuring a full three-dimensional correlation function (discussed below), then the random catalog must also accurately include the redshift selection of the data. The random catalog should also be large enough to not introduce Poisson error in the estimator. This can be checked by ensuring that the RR pair counts in the smallest bin are high enough such that Poisson errors are subdominant.

3 Angular Clustering

The spatial distribution of galaxies can be measured either in two dimensions as projected onto the plane of the sky or in three dimensions using the redshift of each galaxy. As it can be observationally expensive to obtain spectra for large samples of (particularly faint) galaxies, redshift information is not always available for a given sample. One can then measure the two-dimensional projected angular correlation function $\omega(\theta)$ defined as the probability above Poisson of finding two galaxies with an angular separation θ,

$$dP = N[1 + \omega(\theta)]d\Omega, \tag{8.5}$$

where N is the mean number of galaxies per steradian and dΩ is the solid angle of a second galaxy at a separation θ from a randomly chosen galaxy.

Measurements of $\omega(\theta)$ are known to be low by an additive factor known as the "integral constraint," which results from using the data sample itself (which often does not cover large areas of the sky) to estimate the mean galaxy density. This correction becomes important on angular scales comparable to the survey size; on much smaller scales, it is negligible. One can either restrict measurements of the angular clustering to scales where the integral constraint is not important or estimate the amplitude of the integral constraint correction by doubly integrating an assumed power-law form of $\omega(\theta)$ over the survey area, Ω, using

$$\mathrm{AC} = \frac{1}{\Omega}\int\int \omega(\theta)d\Omega_1 d\Omega_2, \tag{8.6}$$

where Ω is the area subtended by the survey. In practice, this can be numerically estimated over the survey geometry using the random catalog itself (see Roche and Eales 1999 for details).

The projected angular two-point correlation function, $\omega(\theta)$, can generally be fit with a power law,

$$\omega(\theta) = A_\omega \theta^\delta, \tag{8.7}$$

where A is the clustering amplitude at a given scale (often $1'$) and δ is the slope of the correlation function.

From measurements of $\omega(\theta)$, one can infer the three-dimensional spatial two-point correlation function, $\xi(r)$, if the redshift distribution of the sources is well known. The two-point correlation function, $\xi(r)$, is usually fit as a power law, $\xi(r) = (r/r_0)^{-\gamma}$, where r_0 is the characteristic scale length of the galaxy clustering defined as the scale at which $\xi = 1$. As the two-dimensional galaxy clustering seen in the plane of the sky is a projection of the three-dimensional clustering, $\omega(\theta)$ is directly related to its three-dimensional analog $\xi(r)$. For a given $\xi(r)$, one can predict the amplitude and slope of $\omega(\theta)$ using Limber's equation, effectively integrating $\xi(r)$ along the redshift direction (e.g., Peebles 1980). If one assumes $\xi(r)$ (and thus $\omega(\theta)$) to be a power law over the redshift range of interest, such that

$$\xi(r,z) = \left[\frac{r_0(z)}{r}\right]^\gamma, \tag{8.8}$$

then

$$w(\theta) = \sqrt{\pi}\frac{\Gamma[(\gamma-1)/2]}{\Gamma(\gamma/2)}\frac{A}{\theta^{\gamma-1}}, \tag{8.9}$$

where Γ is the usual gamma function. The amplitude factor, A, is given by

$$A = \frac{\int_0^\infty r_0^\gamma(z)g(z)\left(\frac{dN}{dz}\right)^2 dz}{\left[\int_0^\infty \left(\frac{dN}{dz}\right)dz\right]^2}, \tag{8.10}$$

where dN/dz is the number of galaxies per unit redshift interval and $g(z)$ depends on γ and the cosmological model,

$$g(z) = \left(\frac{dz}{dr}\right) r^{(1-\gamma)} F(r). \tag{8.11}$$

Here, $F(r)$ is the curvature factor in the Robertson-Walker metric,

$$ds^2 = c^2 dt^2 - a^2 [dr^2/F(r)^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2]. \tag{8.12}$$

If the redshift distribution of sources, dN/dz, is well known, then the amplitude of $\omega(\theta)$ can be predicted for a given power-law model of $\xi(r)$, such that measurements of $\omega(\theta)$ can be used to place constraints on the evolution of $\xi(r)$.

Interpreting angular clustering results can be difficult, however, as there is a degeneracy between the inherent clustering amplitude and the redshift distribution of the sources in the sample. For example, an observed weakly clustered signal projected on the plane of the sky could be due either to the galaxy population being intrinsically weakly clustered and projected over a relatively short distance along the line of sight, or it could result from an inherently strongly clustered distribution integrated over a long distance, which would wash out the signal. The uncertainty on the redshift distribution is therefore often the dominant error in analyses using angular clustering measurements. The assumed galaxy redshift distribution (dN/dz) has varied widely in different studies, such that similar observed angular clustering results have led to widely different conclusions. A further complication is that each sample usually spans a large range of redshifts and is magnitude limited, such that the mean intrinsic luminosity of the galaxies is changing with redshift within a sample, which can hinder interpretation of the evolution of clustering measured in $\omega(\theta)$ studies.

Many of the first measurements of large-scale structure were studies of angular clustering. One of the earliest determinations was the pioneering work of Peebles (1975) using photographic plates from the Lick survey (➋ *Fig. 8-1*). They found $\omega(\theta)$ to be well fit by a power law with a slope of $\delta = -0.8$. Later studies using CCDs were able to reach deeper magnitude limits and found that fainter galaxies had a lower clustering amplitude. One such study was conducted by Postman et al. (1998), who surveyed a contiguous $4^\circ \times 4^\circ$ field to a depth of $I_{\mathrm{AB}} = 24$ mag, reaching to $z \sim 1$. Later surveys that covered multiple fields on the sky found similar results. The lower clustering amplitude observed for galaxies with fainter apparent magnitudes can in principle be due either to clustering being a function of luminosity and/or a function of redshift. To disentangle this dependence, each author assumes a dN/dz distribution for galaxies as a function of apparent magnitude and then fits the observed $\omega(\theta)$ with different models of the redshift dependence of clustering. While many authors measure similar values of the dependence of $\omega(\theta)$ on apparent magnitude, due to differences in the assumed dN/dz distributions, different conclusions are reached regarding the amount of luminosity and redshift dependence to galaxy clustering. Additionally, quoted error bars on the inferred values of r_0 generally include only Poisson and/or cosmic variance error estimates, while the dominant error is often the lack of knowledge of dN/dz for the particular sample in question.

Because of the sensitivity of the inferred value of r_0 on the redshift distribution of sources, it is preferable to measure the three-dimensional correlation function. While it is much easier to interpret three-dimensional clustering measurements, in cases where it is still not feasible to obtain redshifts for a large fraction of galaxies in the sample, angular clustering measurements are still employed. This is currently the case in particular with high redshift and/or very dusty, optically obscured galaxy samples, such as submillimeter galaxies (e.g., Brodwin et al. 2008; Maddox et al. 2010). However, without knowledge of the redshift distribution of the sources, these measurements are hard to interpret.

4 Real and Redshift Space Clustering

Measurements of the two-point correlation function use the redshift of a galaxy, not its distance, to infer its location along the line of sight. This introduces two complications: one is that a cosmological model has to be assumed to convert measured redshifts to inferred distances, and the other is that peculiar velocities introduce redshift space distortions in ξ parallel to the line of sight of Sargent and Turner (1977). On the first point, errors on the assumed cosmology are generally subdominant, so that in theory one could assume different cosmological parameters and check which results are consistent with the assumed values, which is generally not necessary. On the second point, redshift space distortions can be measured to constrain cosmological parameters, and they can also be integrated over to recover the underlying real-space correlation function.

On small spatial scales ($\lesssim 1\,h^{-1}$ Mpc), within collapsed virialized overdensities such as groups and clusters, galaxies have large random motions relative to each other. Therefore, while all of the galaxies in the group or cluster have a similar physical distance from the observer, they have somewhat different redshifts. This causes an elongation in redshift space maps along the line of sight within overdense regions, which is referred to as "Fingers of God." The result is that groups and clusters appear to be radially extended along the line of sight toward the observer. This effect can be seen clearly in ➋ *Fig. 8-4*, where the lower left panel, shows galaxies in redshift space with large "Fingers of God" pointing back to the observer, while in the lower right panel the "Fingers of God" have been modeled and removed. Redshift space distortions are also seen on larger scales ($\gtrsim 1\,h^{-1}$ Mpc) due to streaming motions of galaxies that are infalling onto structures that are still collapsing. Adjacent galaxies will all be moving in the same direction, which leads to coherent motion and causes an apparent *contraction* of structure along the line of sight in redshift space (Kaiser 1987), in the opposite sense as the "Fingers of God."

Redshift space distortions can be clearly seen in measurements of galaxy clustering. While redshift space distortions can be used to uncover information about the underlying matter density and thermal motions of the galaxies (discussed below), they complicate a measurement of the two-point correlation function in real space. Instead of $\xi(r)$, what is measured is $\xi(s)$, where s is the redshift space separation between a pair of galaxies. While some results in the literature present measurements of $\xi(s)$ for various galaxy populations, it is not straightforward to compare results for different galaxy samples and different redshifts as the amplitude of redshift space distortions differs depending on the galaxy type and redshift. Additionally, $\xi(s)$ does not follow a power law over the same scales as $\xi(r)$, as redshift space distortions on both small and large scales decrease the amplitude of clustering relative to intermediate scales.

The real-space correlation function, $\xi(r)$, measures the underlying physical clustering of galaxies independent of any peculiar velocities. Therefore, in order to recover the real-space correlation function, one can measure ξ in two dimensions, both perpendicular to and along the line of sight. Following Fisher et al. (1994), $\mathbf{v}_1$ and $\mathbf{v}_2$ are defined to be the redshift positions of a pair of galaxies, $\mathbf{s}$ to be the redshift space separation $(\mathbf{v}_1 - \mathbf{v}_2)$ and $\mathbf{l} = \frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2)$ to be the mean distance to the pair. The separation between the two galaxies across (r_p) and along (π) the line of sight are defined as

$$\pi = \frac{\mathbf{s} \cdot \mathbf{l}}{|\mathbf{l}|}, \tag{8.13}$$

$$r_p = \sqrt{\mathbf{s} \cdot \mathbf{s} - \pi^2}. \tag{8.14}$$

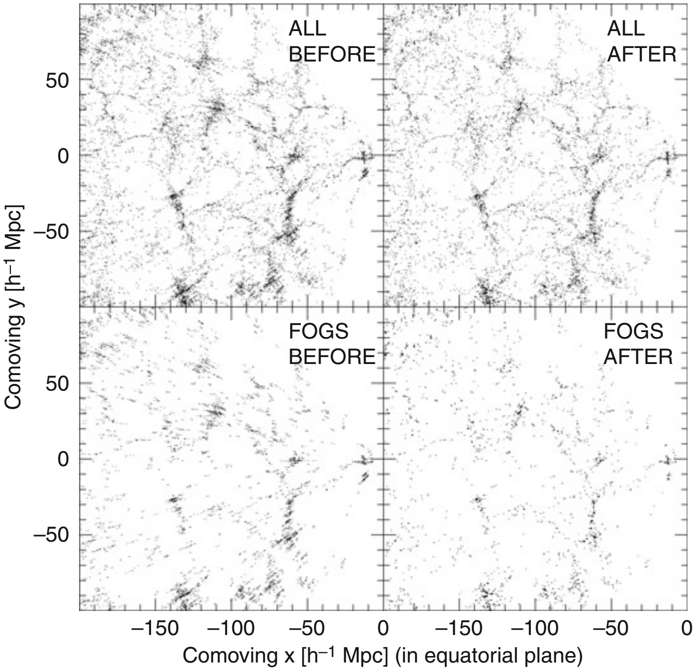


Fig. 8-4

An illustration of the "Fingers of God" (FoG) or elongation of virialized structures along the line of sight, from Tegmark et al. (2004). Shown are galaxies from a slice of the SDSS sample (projected here through the declination direction) in two-dimensional comoving space. The *top row* shows all galaxies in this slice (67,626 galaxies in total), while the *bottom row* shows galaxies that have been identified as having "Fingers of God." The *right column* shows the position of these galaxies in this space after modeling and removing the effects of the "Fingers of God." The observer is located at (x,y=0,0), and the "Fingers of God" effect can be seen in the *lower left panel* as the positions of galaxies being radially smeared along the line of sight toward the observer

One can then compute pair counts over a two-dimensional grid of separations to estimate $\xi(r_p, \pi)$. $\xi(s)$, the one-dimensional redshift space correlation function, is then equivalent to the azimuthal average of $\xi(r_p, \pi)$.

An example of a measurement of $\xi(r_p, \pi)$ is shown in ➋ *Fig. 8-5*. Plotted is ξ as a function of separation r_p (defined in this figure to be σ) across and π along the line of sight. What is usually shown is the upper right quadrant of this figure, which here has been reflected about both axes to emphasize the distortions. Contours of constant ξ follow the color coding, where yellow corresponds to large ξ values and green to low values. On small scales across the line of sight (r_p or $\sigma <\sim 2$ h^{-1} Mpc) the contours are clearly elongated in the π direction; this reflects the "Fingers of God" from galaxies in virialized overdensities. On large scales across the line of sight (r_P or $\sigma >\sim 10$ h^{-1} Mpc) the contours are flattened along the line of sight, due to "the Kaiser effect." This indicates that galaxies on these linear scales are coherently streaming onto structures that are still collapsing.

As this effect is due to the gravitational infall of galaxies onto massive forming structures, the strength of the signature depends on Ω_{matter}. Kaiser (1987) derived that the large-scale anisotropy in the $\xi(r_p, \pi)$ plane depends on $\beta \equiv \Omega_{\text{matter}}/b$ on linear scales, where b is the bias or the ratio of density fluctuations in the galaxy population relative to that of dark matter (discussed further in the next section below). Anisotropies are quantified using the multipole moments of $\xi(r_p, \pi)$ defined as

$$\xi_l(s) = (2l+1)/2 \int \xi(r_p, \pi)\, P_l(\cos\theta)\, d\cos\theta, \tag{8.15}$$

where s is the distance as measured in redshift space, P_l are Legendre polynomials, and θ is the angle between s and the line of sight. The ratio ξ_2/ξ_0, the quadrupole-to-monopole moments of the two-point correlation function, is related to β in a simple manner using linear theory (Hamilton 1998),

$$\xi_2/\xi_0 = f(n)\frac{\frac{4}{3}\beta + \frac{4}{7}\beta^2}{1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2}, \tag{8.16}$$

where $f(n) = (3+n)/n$ and n is the index of the two-point correlation function in a power-law form: $\xi \propto r^{-(3+n)}$ (Hamilton 1992).

Peacock et al. (2001) find using measurements of the quadrupole-to-monopole ratio in the 2dFGRS data (see *Fig. 8-5*) that $\beta = 0.43 \pm 0.07$. For a bias value of around unity (see Sect. 5 below), this implies a low value of $\Omega_{\text{matter}} \sim 0.3$. Similar measurements have been made with

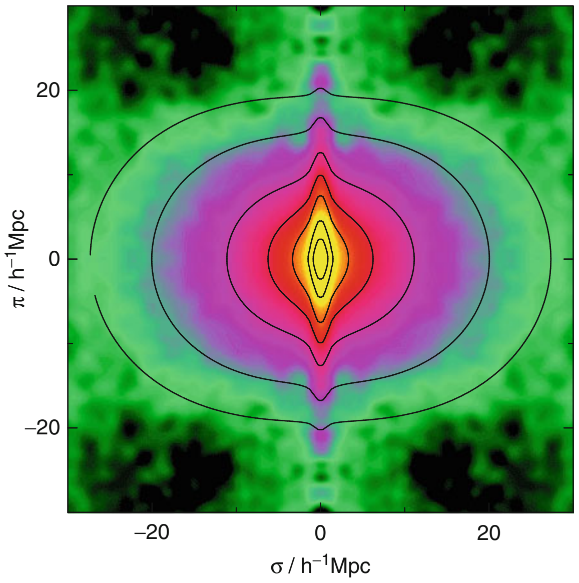


Fig. 8-5
The two-dimensional redshift space correlation function from 2dFGRS (from Peacock et al. 2001). Shown is $\xi(r_p, \pi)$ (in the figure, σ is used instead of r_p), the correlation function as a function of separation across (σ or r_p) and along (π) the line of sight. Contours show lines of constant ξ at ξ =10, 5, 2, 1, 0.5, 0.2, 0.1. Data from the first quadrant (*upper right*) are reflected about both the σ and π axes to emphasize deviations from circular symmetry due to redshift space distortions

clustering measurements using data from the SDSS. Very large galaxy samples are needed to detect this coherent infall and obtain robust estimates of β. At higher redshift, Guzzo et al. (2008) find $\beta = 0.70 \pm 0.26$ at $z = 0.77$ using data from the VVDS and argue that measurements of β as a function of redshift can be used to trace the expansion history of the Universe. We return to the discussion of redshift space distortions on small scales below in ➋ Sect. 6.3.

What is often desired, however, is a measurement of the real-space clustering of galaxies. To recover $\xi(r)$, one can then project $\xi(r_p, \pi)$ along the r_p axis. As redshift space distortions affect only the line-of-sight component of $\xi(r_p, \pi)$, integrating over the π direction leads to a statistic $w_p(r_p)$, which is independent of redshift space distortions. Following Davis and Peebles (1983),

$$w_p(r_p) = 2\int_0^\infty d\pi\, \xi(r_p, \pi) = 2\int_0^\infty dy\, \xi(r_p^2 + y^2)^{1/2}, \tag{8.17}$$

where y is the real-space separation along the line of sight. If $\xi(r)$ is modeled as a power law, $\xi(r) = (r/r_0)^{-\gamma}$, then r_0 and γ can be readily extracted from the projected correlation function, $w_p(r_p)$, using an analytic solution to (➋ 8.17),

$$w_p(r_p) = r_p \left(\frac{r_0}{r_p}\right)^{\gamma} \frac{\Gamma(\frac{1}{2})\Gamma(\frac{\gamma-1}{2})}{\Gamma(\frac{\gamma}{2})}, \tag{8.18}$$

where Γ is the usual gamma function. A power-law fit to $w_p(r_p)$ will then recover r_0 and γ for the real-space correlation function, $\xi(r)$. In practice, (➋ 8.17) is not integrated to infinite separations. Often, values of $\pi_{\max}$ are ~40–80 h^{-1} Mpc, which includes most correlated pairs. It is worth noting that the values of r_0 and γ inferred are covariant. One must therefore be careful when comparing clustering amplitudes of different galaxy populations; simply comparing the r_0 values may be misleading if the correlation function slopes are different. It is often preferred to compare the galaxy bias instead (see next section).

As a final note on measuring the two-point correlation function, as can be seen from ➋ *Fig. 8-3*, flux-limited galaxy samples contain a higher density of galaxies at lower redshift. This is purely an observational artifact due to the apparent magnitude limit including intrinsically lower luminosity galaxies nearby, while only tracing the higher luminosity galaxies further away. As discussed below in ➋ Sect. 6, because the clustering amplitude of galaxies depends on their properties, including luminosity, one would ideally only measure $\xi(r)$ in volume-limited samples where galaxies of the same absolute magnitude are observed throughout the entire volume of the sample, including at the highest redshifts. Therefore, often the full observed galaxy population is not used in measurements of $\xi(r)$, rather volume-limited subsamples are created where all galaxies are brighter than a given absolute magnitude limit. This greatly facilitates the theoretical interpretation of clustering measurements (see ➋ Sect. 8) and the comparison of results from different surveys.

5 Galaxy Bias

It was realized decades ago that the spatial clustering of observable galaxies need not precisely mirror the clustering of the bulk of the matter in the Universe. In its most general form, the galaxy density can be a nonlocal and stochastic function of the underlying dark matter density. This galaxy "bias" – the relationship between the spatial distribution of galaxies and the underlying dark matter density field – is a result of the varied physics of galaxy formation which can cause the spatial distribution of baryons to differ from that of dark matter. Stochasticity appears to have little effect on bias except for adding extra variance (e.g., Scoccimarro 2000), and nonlocality can be taken into account to first order by using smoothed densities over larger scales. In this approximation, the smoothed galaxy density contrast is a general function of the underlying dark matter density contrast on some scale,

$$\delta_g = f(\delta), \tag{8.19}$$

where $\delta \equiv (\rho/\bar{\rho}) - 1$ and $\bar{\rho}$ is the mean mass density on that scale. If we assume $f(\delta)$ is a linear function of δ, then we can define the linear galaxies bias b as the ratio of the mean overdensity of galaxies to the mean overdensity of mass,

$$b = \delta_g/\delta, \tag{8.20}$$

and can in theory depend on scale and galaxy properties such as luminosity, morphology, color, and redshift. In terms of the correlation function, the linear bias is defined as the square root of the ratio of the two-point correlation function of the galaxies relative to the dark matter

$$b = (\xi_{\mathrm{gal}}/\xi_{\mathrm{dark\ matter}})^{1/2} \tag{8.21}$$

and is a function of scale. Note that $\xi_{\mathrm{dark\ matter}}$ is the Fourier transform of the dark matter power spectrum. The bias of galaxies relative to dark matter is often referred to as the absolute bias as opposed to the relative bias between galaxy populations (discussed below).

The concept of galaxies being a biased tracer of the underlying total mass field (which is dominated by dark matter) was introduced by Kaiser (1984) in an attempt to reconcile the different clustering scale lengths of galaxies and rich clusters, which could not both be unbiased tracers of mass. Kaiser (1984) shows that clusters of galaxies would naturally have a large bias as a result of being rare objects which formed at the highest density peaks of the mass distribution above some critical threshold. This idea is further developed analytically by Bardeen et al. (1986) for galaxies, who show that for a Gaussian distribution of initial mass density fluctuations, the peaks which first collapse to form galaxies will be more clustered than the underlying mass distribution. Mo and White (1996) use extended Press-Schechter theory to determine that the bias depends on the mass of the dark matter halo as well as the epoch of galaxy formation and that a linear bias is a decent approximation well into the nonlinear regime, where $\delta > 1$. The evolution of bias with redshift is developed in theoretical work by Fry (1996) and Tegmark and Peebles (1998), who find that the bias is naturally larger at earlier epochs of galaxy formation, as the first galaxies to form will collapse in the most overdense regions of space, which are biased (akin to mountain peaks being clustered). They further show that regardless of the initial amplitude of the bias factor, with time galaxies, will become unbiased tracers of the mass distribution ($b \to 1$ as $t \to \infty$). Additionally, Mann et al. (1998) find that while bias is generally scale dependent, the dependence is weak, and on large scales, the bias tends toward a constant value.

A galaxy population can be "anti-biased" if $b < 1$, indicating that galaxies are less clustered than the dark matter distribution. As discussed below, this appears to be the case for some galaxy samples at low redshift. The galaxy bias of a given observational sample is often inferred by comparing the observed clustering of galaxies with the clustering of dark matter measured in a cosmological simulation. Therefore, the bias depends on the cosmological model used in the simulation. The dominant relevant cosmological parameter is σ_8 defined as the standard deviation of galaxy count fluctuations in a sphere of radius $8\,\mathrm{h}^{-1}$ Mpc, and the absolute bias value inferred can be simply scaled with the assumed value of σ_8. As discussed in ➋ Sect. 9.1 below, the absolute galaxy bias can also be estimated from the data directly, without having to resort to comparisons with cosmological simulations, by using the ratio of the two-point and three-point correlation functions, which have different dependencies on the bias. While this measurement can be somewhat noisy, it has the advantage of not assuming a cosmological model from which to derive the dark matter clustering. This measurement is performed by Verde et al. (2002) and Gaztañaga et al. (2005), who find that galaxies in 2dFGRS have a linear bias value very close to unity on large scales.

The relative bias between different galaxy populations can also be measured and is defined as the ratio of the clustering of one population relative to another. This is often measured using the ratio of the projected correlation functions of each population,

$$b_{\text{gal1/gal2}} = (w_{p,\text{gal1}}/w_{p,\text{gal2}})^{1/2}, \tag{8.22}$$

where both measurements of $w_p(r_p)$ have been integrated to the same value of π_{max}. The relative bias is used to compare the clustering of galaxies as a function of observed parameters and does not refer to the clustering of dark matter. It is a useful way to compare the observed clustering for different galaxy populations without having to rely on an assumed value of σ_8 for the dark matter.

6 The Dependence of Clustering on Galaxy Properties

The two-point correlation function has long been known to depend on galaxy properties and can vary as a function of galaxy luminosity, morphological or spectral type, color, stellar mass, and redshift. The general trend is that galaxies that are more luminous, early-type, bulge-dominated, optically red, and/or higher stellar mass are more clustered than galaxies that are less luminous, late type, disk-dominated, optically blue, and/or lower stellar mass. Presented below are relatively recent results indicating how clustering properties depend on galaxy properties from the largest redshifts surveys currently available. The physical interpretation behind these trends is presented in ➋ Sect. 8 below.

6.1 Luminosity Dependence

➋ *Figure 8-3* shows the large-scale structure reflected in the galaxy distribution at low redshift. What is plotted is the spatial distribution of galaxies in a flux-limited sample, meaning that all galaxies down to a given apparent magnitude limit are included. This results in the apparent lack of galaxies or structure at higher redshift in the figure as at large distances only the most luminous galaxies will be included in a flux-limited sample. In order to robustly determine the underlying clustering, one should, if possible, create volume-limited subsamples in which galaxies of the same luminosity can be detected at all redshifts. In this way, the mean luminosity of the sample does not change with redshift and galaxies at all redshifts are weighted equally.

The left panel of ➋ *Fig. 8-6* shows the projected correlation function, $w_p(r_p)$, for galaxies in SDSS in volume-limited subsamples corresponding to different absolute magnitude ranges. The more luminous galaxies are more strongly clustered across a wide range in absolute optical magnitude, from $-17 < M_r < -23$. Power-law fits on scales from ~0.1 to ~10 h^{-1} Mpc show that while the clustering amplitude depends sensitively on luminosity, the slope does not. Only in

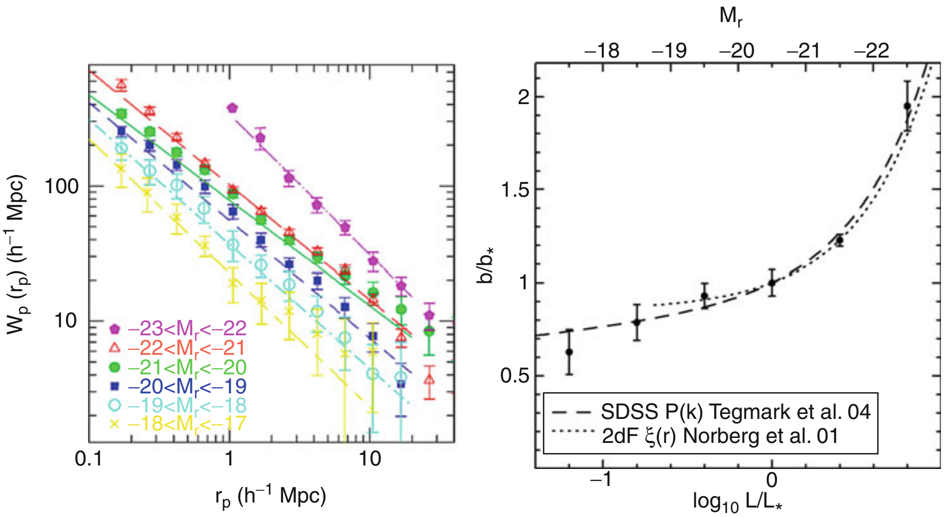


Fig. 8-6

Luminosity dependence of galaxy clustering. On the *left* is shown the projected correlation function, $w_p(r_p)$, for SDSS galaxies in different absolute magnitude ranges where brighter galaxies are seen to be more clustered. On the right is the relative bias of galaxies as a function of luminosity (Both figures are from Zehavi et al. (2005))

the brightest and faintest magnitude bins does the slope deviate from $\gamma \sim 1.8$ and have a steeper value of $\gamma \sim 2.0$. Across this magnitude, range r_0 varies from ~2.8 h^{-1} Mpc at the faint end to ~10 h^{-1} Mpc at the bright end. This same general trend is found in the 2dFGRS and other redshift surveys (e.g., Norberg et al. 2001).

The right panel of *Fig. 8-6* shows the relative bias of SDSS galaxies as a function of luminosity, relative to the clustering of L^* galaxies, measured at the scale of $r_p = 2.7\,h^{-1}$ Mpc, which is in the nonlinear regime, where $\delta > 1$ (Zehavi et al. 2005). L^* is the characteristic galaxy luminosity defined as the luminosity of the break in the galaxy luminosity function. The relative bias is seen to steadily increase at higher luminosity and rise sharply above L^*. This is in good agreement with the results from Tegmark et al. (2004), using the power spectrum of SDSS galaxies measured in the linear regime on a scale of ~100 h^{-1} Mpc. The data also agree with the clustering results of galaxies in the 2dFGRS from Norberg et al. (2001). The overall shape of the relative bias with luminosity indicates a slow rise up to the value at L^*, above which the rise is much steeper. As discussed in Sect. 8.2 below, this trend shows that brighter galaxies reside in more massive dark matter halos than fainter galaxies.

6.2 Color and Spectral-Type Dependence

The clustering strength of galaxies also depends on restframe color and spectral type, with a stronger dependence than on luminosity. *Figure 8-7* shows the spatial distribution of galaxies in SDSS, color coded as a function of restframe color. Red galaxies are seen to preferentially populate the most overdense regions, while blue galaxies are more smoothly distributed in space.

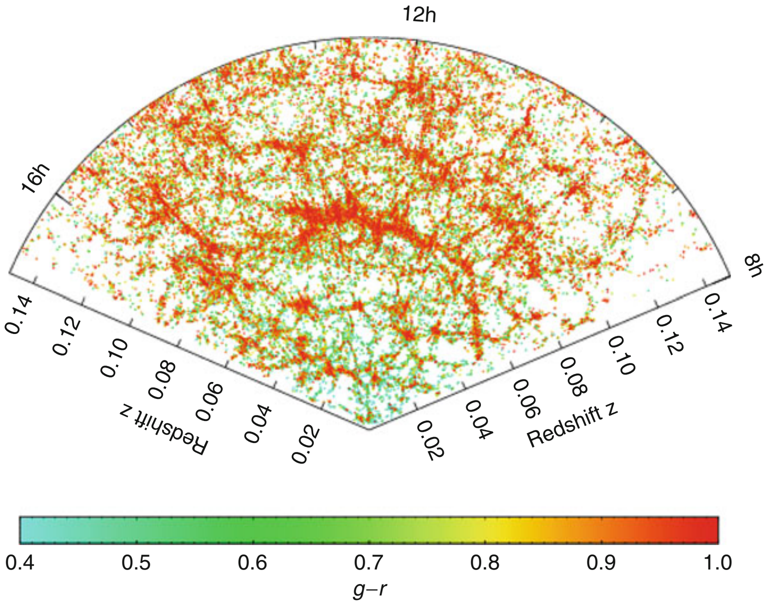


Fig. 8-7
The spatial distribution of galaxies in the SDSS main galaxy sample as a function of redshift and right ascension projected through 8° in declination and color-coded by restframe optical color. *Red* galaxies are seen to be more clustered than *blue* galaxies and generally trace the centers of groups and clusters, while *blue* galaxies populate further into the galaxy voids (Taken from Zehavi et al. (2011))

This is reflected in the correlation function of galaxies split by restframe color. Red galaxies have a larger correlation length and steeper slope than blue galaxies: $r_0 \sim 5–6\ \mathrm{h}^{-1}$ Mpc and $\gamma \sim 2.0$ for red L^* galaxies, while $r_0 \sim 3–4\ \mathrm{h}^{-1}$ Mpc and $\gamma \sim 1.7$ for blue L^* galaxies in SDSS (Zehavi et al. 2005). Clustering studies from the 2dFGRS split the galaxy sample at low redshift by spectral type into galaxies with emission line spectra versus absorption line spectra, corresponding to star forming and quiescent galaxies, and find similar results: that quiescent galaxies have larger correlation lengths and steeper clustering slopes than star-forming galaxies (Madgwick et al. 2003).

Red and blue galaxies have distinct luminosity-dependent clustering properties. As shown in *Fig. 8-8*, the general trends seen in r_0 and γ with luminosity for all galaxies are well reflected in the blue galaxy population; however, at faint luminosities ($L \lesssim 0.5L^*$), red galaxies have larger clustering amplitudes and slopes than L^* red galaxies. This reflects the fact that faint red galaxies are often found distributed throughout galaxy clusters.

Galaxy clustering also depends on other galaxy properties such as stellar mass, concentration index, and the strength of the 4,000Å break ($D_{4,000}$) in that galaxies that have larger stellar mass, more centrally concentrated light profiles, and/or larger $D_{4,000}$ measurements (indicating older stellar populations) are more clustered (Li et al. 2006). This is not surprising given the observed trends with luminosity and color and the known dependencies of other galaxy properties with luminosity and color. Clearly, the galaxy bias is a complicated function of various galaxy properties.

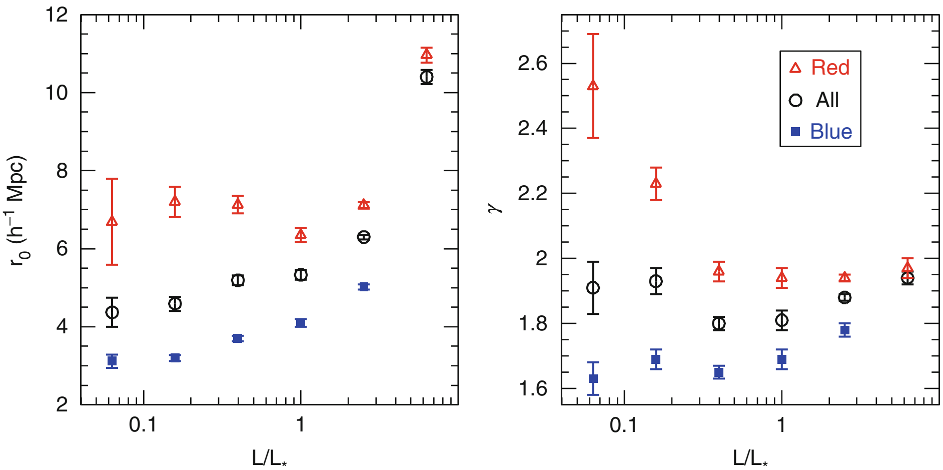


Fig. 8-8

The clustering scale length, r_0 (*left*), and slope, γ (*right*), for all, *red* and *blue*, galaxies in SDSS as a function of luminosity. While all galaxies are more clustered at brighter luminosities and *red* galaxies are more clustered than blue galaxies at all luminosities, below L^*, the *red* galaxy clustering length increases at fainter luminosities. The clustering slope for faint *red* galaxies is also much steeper than at other luminosities (Taken from Zehavi et al. (2011))

6.3 Redshift Space Distortions

The fact that red galaxies are more clustered than blue galaxies is related to the morphology-density relation (Dressler 1980), which results from the fact that galaxies with elliptical morphologies are more likely to be found in regions of space with a higher local surface density of galaxies. The redshift space distortions seen for red and blue galaxies also show this.

As discussed in Sect. 4, redshift space distortions arise from two different phenomena: virialized motions of galaxies within collapsed overdensities such as groups and clusters and the coherent streaming motion of galaxies onto larger structures that are still collapsing. The former is seen on relatively small scales ($r_p \lesssim 1\,h^{-1}$ Mpc) while the latter is detected on larger scales ($r_P \gtrsim 1\,h^{-1}$ Mpc). While the presence of redshift space distortions complicates the measurement of the real space $\xi(r)$, these distortions can be used to uncover information about the thermal motions of galaxies in groups and clusters as well as the amplitude of the mass density of the Universe, Ω_{matter}.

Figure 8-9 shows $\xi(r_p, \pi)$ for quiescent and star-forming galaxies in 2dF. The quiescent galaxies on the left show larger "Fingers of God" than the star-forming galaxies on the right, reflecting the fact that red, quiescent galaxies have larger motions relative to each others. This naturally arises if red, quiescent galaxies reside in more massive, virialized overdensities with larger random peculiar velocities than star-forming, optically blue galaxies. The large-scale coherent infall of galaxies is seen both for blue and red galaxies, though it is often easier to see for blue galaxies, due to their smaller "Fingers of God."

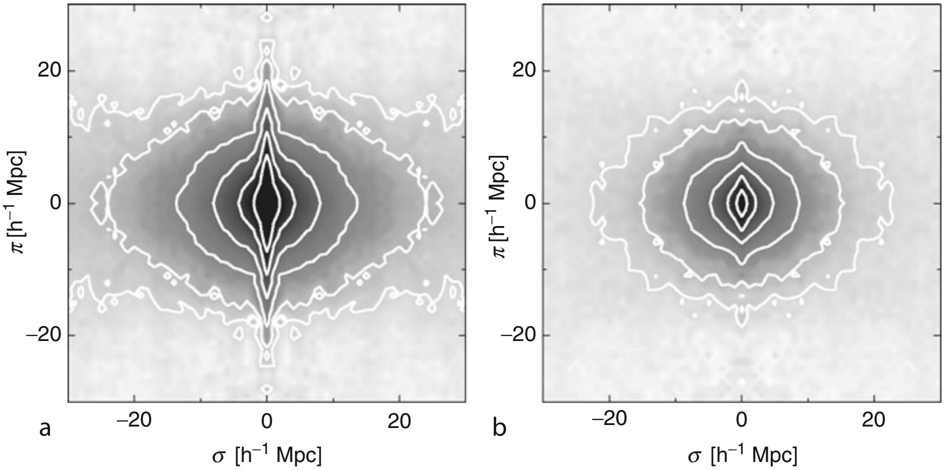


Fig. 8-9
Two-dimensional redshift space correlation function $\xi(r_p, \pi)$ (as in *Fig. 8-6*, here, σ is used instead of r_p) for quiescent, absorption line galaxies on the *left* and star-forming, emission line galaxies on the *right*. The redshift space distortions are different for the different galaxy populations, with quiescent and/or *red* galaxies showing more pronounced "Fingers of God." Both galaxy types exhibit coherent infall on large scales. Contours show lines of constant ξ at ξ =10, 5, 2, 1, 0.5, 0.2, 0.1 (Taken from Madgwick et al. (2003))

These small-scale redshift space distortions can be quantified using the σ_{12} statistic known as the pairwise velocity dispersion (Davis et al. 1978; Fisher et al. 1994). This is measured by modeling $\xi(r_p, \pi)$ in real space, which is then convolved with a distribution of random pairwise motions, $f(v)$, such that

$$\xi(r_p, \pi) = \int_{-\infty}^{\infty} \xi'(r_p, \pi - v/H_0) f(v) dv, \tag{8.23}$$

where the random motions are often taken to have an exponential form, which has been found to fit observed data well:

$$f(v) = \frac{1}{\sigma_{12}\sqrt{2}} \exp\left(-\frac{\sqrt{2}|v|}{\sigma_{12}}\right). \tag{8.24}$$

In the 2dFGRS, Madgwick et al. (2003) find that $\sigma_{12} = 416 \pm 76$ km s^{-1} for star-forming galaxies and $\sigma_{12} = 612 \pm 92$ km s^{-1} for quiescent galaxies measured on scales of 8–20 h^{-1} Mpc. Using SDSS data, Zehavi et al. (2002) find that σ_{12} is ~300–450 km s^{-1} for blue, star-forming galaxies and ~650–750 km s^{-1} for red, quiescent galaxies. It has been shown, however, that this statistic can be sensitive to large, rare overdensities, such that samples covering large volumes are needed to measures it robustly.

Madgwick et al. (2003) further measure the large-scale anisotropies seen in $\xi(r_p, \pi)$ for galaxies split by spectral type and find that $\beta = 0.49 \pm 0.13$ for star-forming galaxies and $\beta = 0.48 \pm 0.14$ for quiescent galaxies. This implies a similar bias for both galaxy types on large scales, though they find that on smaller scales integrated up to 8 h^{-1} Mpc, the relative bias of quiescent to star-forming galaxies is $b_{\text{rel}} = 1.45 \pm 0.14$.

7 The Evolution of Galaxy Clustering

The observed clustering of galaxies is expected to evolve with time as structure continues to grow due to gravity. The exact evolution depends on cosmological parameters such as Λ and Ω_{matter}. Larger values of Λ, for example, lead to larger voids and higher density contrasts between overdense and underdense regions. By measuring the clustering of galaxies at higher redshift, one can break degeneracies that exist between the galaxy bias and cosmological parameters that are constrained by low redshift clustering measurements. It is therefore useful to determine the clustering of galaxies as a function of cosmic epoch not only to further constrain cosmological parameters but also galaxy evolution.

One might expect the galaxy clustering amplitude r_0 to increase over time as overdense regions become more overdense as galaxies move toward groups and clusters due to gravity. However, the exact evolution of the clustering of galaxies depends not only on gravity but also on the expansion history of the Universe and therefore cosmological parameters such as Λ. Additionally, over time, new galaxies form while existing galaxies grow in both mass and luminosity. Therefore, the expected changes of galaxy clustering as a function of redshift depend both on relatively well-known cosmological parameters and more unknown galaxy formation and evolution physics which likely depends on gas accretion, star formation, and feedback processes, as well as mergers.

For a given cosmological model, one can predict how the clustering of dark matter should evolve with time using cosmological N-body simulations. For a ΛCDM Universe, r_0 for dark matter particles is expected to increase from ~0.8 h^{-1} Mpc at $z = 3$ to ~5 h^{-1} Mpc at $z = 0$ (Weinberg et al. 2004). However, according to hierarchical structure formation theories, at high redshifts, the first galaxies to form will be the first structures to collapse, which will be biased tracers of the mass. The galaxy bias is expected to be a strong function of redshift, initially >1 at high redshift and approaching unity over time. Therefore, r_0 for galaxies may be a much weaker function of time than it is for dark matter as the same galaxies are not observed as a function of redshift, and over time, new galaxies form in less biased locations in the Universe.

The projected angular and three-dimensional correlation functions of galaxies have been observed to $z \sim 5$. Star-forming Lyman break galaxies at $z \sim 3$–5 are found to have $r_0 \sim 4$–6 h^{-1} Mpc, with a bias relative to dark matter of ~3–4 (Adelberger et al. 2005; Ouchi et al. 2004). Brighter Lyman break galaxies are found to cluster more strongly than fainter Lyman break galaxies. The correlation length, r_0, is found to be roughly constant between $z = 5$ and $z = 3$, implying that the bias is increasing at earlier cosmic epoch. Spectroscopic galaxy surveys at $z > 2$ are currently limited to samples of at most a few thousand galaxies, so most clustering measurements are angular at these epochs. In one such study by Wake et al. (2011), photometric redshifts of tens of thousands of galaxies at $1 < z < 2$ are used to measure the angular clustering as a function of stellar mass. They find a strong dependence of clustering amplitude on stellar mass in each of three redshift intervals in this range.

At $z \sim 1$, larger spectroscopic galaxy samples exist, and three-dimensional two-point clustering analyses have been performed as a function of luminosity, color, stellar mass, and spectral type. The same general clustering trends with galaxy property that are observed at $z \sim 0$ are also seen at $z \sim 1$, in that galaxies that are brighter, redder, early spectral type, and/or more massive are more clustered. The clustering scale length of red galaxies is found to be ~5–6 h^{-1} Mpc, while for blue galaxies, it is ~3.5–4.5 h^{-1} Mpc, depending on luminosity (Coil et al. 2008; Meneux et al. 2006). At a given luminosity, the observed correlation length is only ~15% smaller at $z = 1$ than $z = 0$, indicating that unlike for dark matter, the galaxy r_0 is roughly constant over time. These results are consistent with predictions from ΛCDM simulations.

The measured values of r_0 at $z \sim 1$ imply that galaxies are more biased at $z = 1$ than at $z = 0$. Within the DEEP2 sample, the bias measured on scales of ~1–10 h^{-1} Mpc varies from ~1.25–1.55, with brighter galaxy samples being more biased tracers of the dark matter (Coil et al. 2006). These results are consistent with the idea that galaxies formed early on in the most overdense regions of space, which are biased.

As in the nearby Universe, the clustering amplitude is a stronger function of color than of luminosity at $z \sim 1$. Additionally, the color-density relation is found to already be in place at $z = 1$, in that the relative bias of red to blue galaxies is as high at $z = 1$ as at $z = 0.1$ (Coil et al. 2008). This implies that the color-density relation is not due to cluster-specific physics, as most galaxies at $z = 1$ in field spectroscopic surveys are not in clusters. Therefore, it must be physical processes at play in galaxy groups that initially set the color and morphology-density relations. Red galaxies show larger "Fingers of God" in $\xi(r_p, \pi)$ measurements than blue galaxies do, again showing that red galaxies at $z = 1$ lie preferentially in virialized, more massive overdensities compared to blue galaxies. Both red and blue galaxies show coherent infall on large scales.

8 Halo Model Interpretation of $\xi(r)$

The current paradigm of galaxy formation posits that galaxies form in the center of larger dark matter halos, collapsed overdensities in the dark matter distribution with $\rho/\bar{\rho} \sim 200$, inside of which all mass is gravitationally bound. The clustering of galaxies can then be understood as a combination of the clustering of dark matter halos, which depends on cosmological parameters, and how galaxies populate dark matter halos, which depends on galaxy formation and evolution physics. For a given cosmological model, the properties of dark matter halos, including their evolution with time, can be studied in detail using N-body simulations. The masses and spatial distribution of dark matter halos should depend only on the properties of dark matter, not baryonic matter, and the expansion history of the Universe; therefore, the clustering of dark matter halos should be insensitive to baryon physics. However, the efficiency of galaxy formation is very dependent on the complicated baryonic physics of, for example, star formation, gas cooling, and feedback processes. The halo model allows the relatively simple cosmological dependence of galaxy clustering to be cleanly separated from the more complex baryonic astrophysics, and it shows how clustering measurements for a range of galaxy types can be used to constrain galaxy evolution physics.

8.1 Estimating the Mean Halo Mass from the Bias

One can use the observed large-scale clustering amplitude of different observed galaxy populations to identify the typical mass of their parent dark matter halos in order to place these galaxies in a cosmological context (see ➲ Chap. 10). The large-scale clustering amplitude of dark matter halos as a function of halo mass is well determined in N-body simulations, and analytic fitting formula are provided by, for example, Mo and White (1996) and Sheth et al. (2001). Analytic models can then predict the clustering of both dark matter particles and galaxies as a function of scale by using the clustering of dark matter halos and the radial density profile of dark matter and galaxies within those halos (Ma and Fry 2000; Peacock and Smith 2000; Seljak 2000). In this scheme, on large, linear scales where $\delta < 1$ ($\rho/\bar{\rho} \sim 1$), the clustering of a given galaxy population can be used to determine the mean mass of the dark matter halos hosting those galaxies for a given cosmological model. To achieve this, the large-scale bias is estimated by comparing the observed galaxy clustering amplitude with that of dark matter in an N-body simulation, and then galaxies are assumed to reside in halos of a given mass that have the same bias in simulations.

Simulations show that higher mass halos cluster more strongly than lower mass halos (Sheth and Tormen 1999). This then leads to an interpretation of galaxy clustering as a function of luminosity in which luminous galaxies reside in more massive dark matter halos than less luminous galaxies. Similarly, red galaxies typically reside in more massive halos than blue galaxies of the same luminosity; this is observationally verified by the larger "Fingers of God" observed for red galaxies. Combining the large-scale bias with the observed galaxy number density further allows one to constrain the fraction of halos that host a given galaxy type by comparing the galaxy space density to the parent dark matter halo space density. This constrains the duty cycle or fraction of halos hosting galaxies of a given population.

8.2 Halo Occupation Distribution Modeling

While such estimates of the mean host halo mass and duty cycle are fairly straightforward to carry out, a greater understanding of the relation between galaxy light and dark matter mass is gleaned by performing halo occupation distribution modeling.

The general halo-based model discussed above, in which the clustering of galaxies reflects the clustering of halos, was further developed by Peacock and Smith (2000) to include the efficiency of galaxy formation, or how galaxies populate halos. The proposed model depends on both the halo occupation number, equal to the number of galaxies in a halo of a given mass, for a galaxy sample brighter than some limit, and the location of the galaxies within these halos. In the Peacock and Smith (2000) model, it is assumed that one galaxy is at the center of the halo (the "central" galaxy), and the rest of the galaxies in the same halo are "satellite" galaxies that trace the dark matter radial mass distribution, which follows an NFW profile (Navarro et al. 1997). The latter assumption results in a general power-law shape for the galaxy correlation function.

A similar idea was proposed by Benson et al. (2000), who used a semi-analytic model in conjunction with a cosmological N-body simulation to show that the observed galaxy $\xi(r)$ could be reproduced with a ΛCDM simulation (though not with a τ CDM simulation with $\Omega_{\text{matter}} = 1$). They also employ a method for locating galaxies inside dark matter halos such that one galaxy resides at the center of all halos above a given mass threshold, while additional galaxies are assigned the location of a random dark matter particle within the same halo, such that galaxies have the same NFW radial profile within halos as the dark matter particles (see ➲ *Fig. 8-10*, left panel).

In these models, the clustering of galaxies on scales larger than a typical halo (~1–2 h^{-1} Mpc) results from pairs of galaxies in separate halos called the "two-halo term," while the clustering on smaller scales ($\lesssim 1\, h^{-1}$ Mpc) is due to pairs of galaxies within the same parent halo called the "one-halo term." When the pairs from these two terms are added together, the resulting galaxy correlation function should roughly follow a power law.

Benson et al. (2000) show that on large scales, there is a simple relation in the bias between galaxies and dark matter halos, while on small scales, the correlation function depends on the number of galaxies in a halo and the finite size of halos. When the clustering signal from these

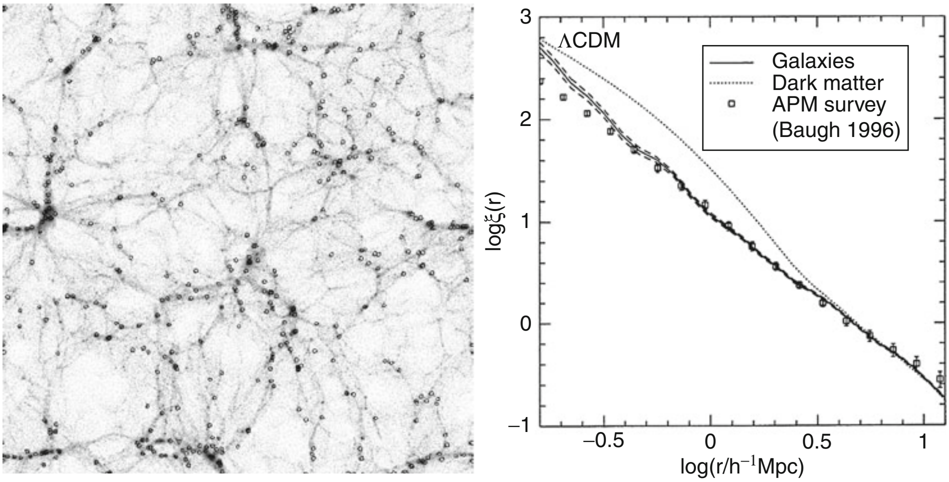


Fig. 8-10

***Left*: The large-scale structure seen in a ΛCDM N-body dark matter only simulation of size 141 × 141 × 8 h^{-3} Mpc^3. The *gray* scale indicates the density of dark matter, while the locations of galaxies are shown with *open circles*. Galaxies are added to the simulation output using a semi-analytic model which assumes that dark matter halos above a given mass threshold have at least one "central" galaxy located at the center of the halo. Higher mass halos contain additional "satellite" galaxies, which are assigned the location of a random dark matter particle in the halo (Taken from Benson et al. (2000)). *Right*: The two-point correlation function of dark matter particles (*dotted line*) and galaxies (*solid line* with *dashed line* showing Poisson error bars) in the simulation of Benson et al. (2000) compared with the observed clustering of galaxies in the APM survey (*open squares*) (Baugh 1996)**

two scales (corresponding to the "two-halo" and "one-halo" terms) is combined, a power law results for the galaxy $\xi(r)$ (right panel, *Fig. 8-10*). Galaxies are found to be anti-biased relative to dark matter (i.e., less clustered than the dark matter) on scales smaller than the typical halo, though the bias is close to unity on larger scales. The clustering of galaxies that results from this semi-analytic model is also found to match the observed clustering of galaxies in the APM survey above a given luminosity threshold (Baugh 1996).

By defining the halo occupation distribution (HOD) as the probability that a halo of a given mass contains N galaxies, $P(N|M)$, Berlind and Weinberg (2002) quantify how the observed galaxy $\xi(r)$ depends on different HOD model parameters. Using N-body simulations, they identify dark matter halos and place galaxies into the simulation using a simple HOD model with two parameters: a minimum mass at which a halo hosts, on average, one central galaxy ($M_{\min}$) at the center of the halo, and the slope (α) of the $P(N|M)$ function for satellite galaxies. The latter determines how many satellite galaxies there are as a function of halo mass. They further assume that the satellite galaxies follow an NFW profile, as the dark matter does, though the concentration of the radial profile can be changed. They show that the "two-halo term" is simply the halo center correlation function weighted by a large-scale bias factor, while the "one-halo term" is sensitive to both α and the concentration of the galaxy profile within halos. Obtaining a power law $\xi(r)$ therefore strongly constrains the HOD model parameters.

Kravtsov et al. (2004) propose that the locations of satellite galaxies within dark matter halos should correspond to locations of subhalos, distinct gravitationally bound regions within the larger dark matter halos, instead of tracing random dark matter particles. Using cosmological N-body simulations, they show that at $z > 1$ $\xi(r)$ for galaxies should deviate strongly from a power law on small scales, due to a rise in the "one-halo term." In this model, the clustering of galaxies can be understood as the clustering of dark matter parent halos and subhalos, and the power-law shape that is observed at $z \sim 0$ is a coincidence of the one- and two-halo terms having similar amplitudes and slopes at the typical scale of halos. They find that the formation and evolution of halos and subhalos through merging and dynamical processes are the main physical drivers of large-scale structure.

With the unprecedentedly large galaxy sample with spectroscopic redshifts that is provided by SDSS, departures from a power law $\xi(r)$ were detected by Zehavi et al. (2004) using a volume-limited subsample of 22,000 galaxies from a parent sample of 118,000 galaxies. The deviations from a power law are small enough at $z \sim 0$ that a large sample covering a sufficiently large cosmological volume is required to overcome the errors due to cosmic variance to detect these small deviations. It is found that there is a change in the slope of $\xi(r)$ on scales of ~1–2 h^{-1} Mpc; this corresponds to the scale at which the one- and two-halo term are equal (see ➋ *Fig. 8-11*). Zehavi et al. (2004) find that $\xi(r)$ measured from the SDSS data is better fit by an HOD model, which includes small deviations from a power law, than by a pure power law. The HOD model that is fit has three parameters: the minimum mass to host a single central galaxy ($M_{\rm min}$), the minimum mass to host a single satellite galaxy (M_1), and the slope of $P(N|M)$ (α), which

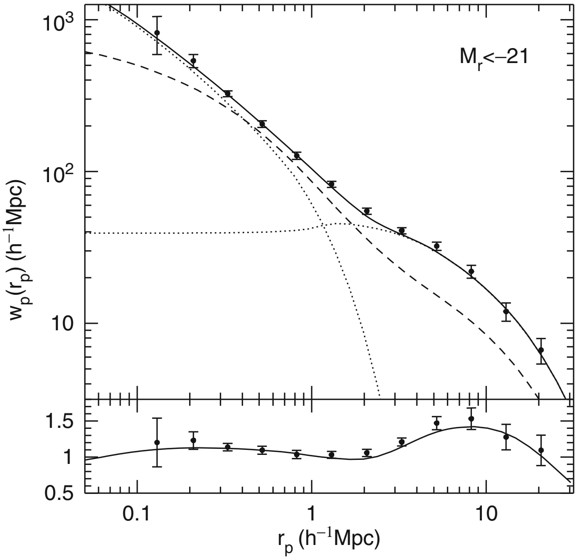


Fig. 8-11

The projected correlation function, $w_p(r_p)$, for SDSS galaxies with $M_r < -21$ is shown as data points with error bars. The best-fit HOD model is shown as a *solid line*, with the contributions from the one and two-halo terms shown with *dotted lines*. The projected correlation function of dark matter at this redshift is shown with a *dashed line*. The *bottom panel* shows deviations in $w_p(r_p)$ for the data and the HOD model from the best-fit power law (Taken from Zehavi et al. (2004))

determines the average number of satellite galaxies as a function of host halo mass. In this model, dark matter halos with $M_{\rm min} < M < M_1$ host a single galaxy, while above M_1, they host, on average, $(M/M_1)^\alpha$ galaxies. Using $w_p(r_p)$, one can fit for M_1 and α, while the observed space density of galaxies is used to derive $M_{\rm min}$. For a galaxy sample with $M_r < -21$, the best-fit HOD parameters are $M_{\rm min} = 6.1 \times 10^{12}\ {\rm h}^{-1} M_\odot$, $M_1 = 4.7 \times 10^{13}\ {\rm h}^{-1} M_\odot$, and $\alpha = 0.89$.

8.3 Interpreting the Luminosity and Color Dependence of Galaxy Clustering

In general, these HOD parameters reflect the efficiency of galaxy formation and evolution and can be a function of galaxy properties such as luminosity, color, stellar mass, and morphology. Zehavi et al. (2011) present HOD fits to SDSS samples as a function of luminosity and color and find that α is generally ~1.0–1.1, though it is a bit higher for the brightest galaxies (~1.3 for $M_r < -22.0$). There is a strong trend between luminosity and halo mass; $M_{\rm min}$ varies as a function of luminosity from ~$10^{11}\ {\rm h}^{-1} M_\odot$ for $M_r < -18$ to ~$10^{14}\ {\rm h}^{-1} M_\odot$ for $M_r < -22$. M_1 is generally ~17 times higher than the value of $M_{\rm min}$ for all luminosity threshold samples (see *Fig. 8-12*). This implies that a halo with two galaxies above a given luminosity is ~17 times more massive than a halo hosting one galaxy above the same luminosity limit. Further, the fraction of galaxies that are satellites decreases at higher luminosities, from ~33% at $M_r < -18$ to 4% at $M_r < -22$. The right panel of *Fig. 8-12* shows the mass-to-light ratio of the virial halo mass to the central galaxy r-band luminosity as a function of halo mass. This figure shows that halos of mass ~ $4 \times 10^{11}\ {\rm h}^{-1} M_\odot$ are maximally efficient at galaxy formation at converting baryons into light.

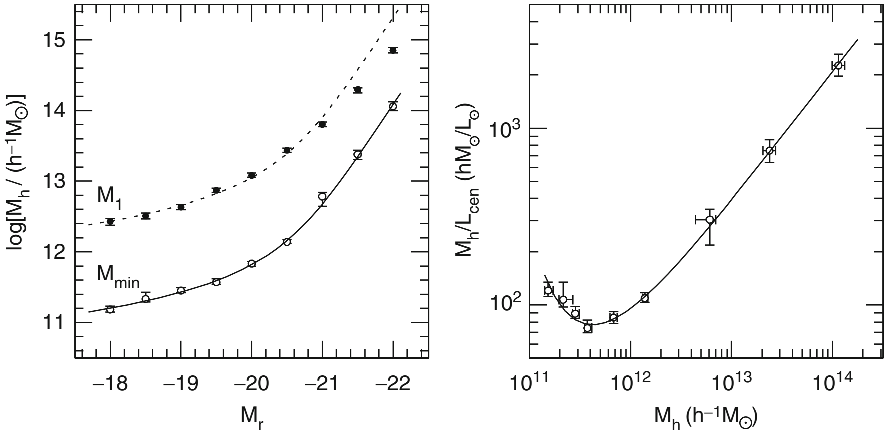


Fig. 8-12
***Left*: The characteristic mass scale of dark matter halos hosting galaxies as a function of the luminosity threshold of the galaxy sample. Both the minimum halo mass to host a single galaxy is shown ($M_{\rm min}$) as well as the minimum mass to host additional satellite galaxies (M_1). A strong relationship clearly exists between halo mass and galaxy luminosity. *Right*: The ratio of the halo mass to the median central galaxy luminosity as a function of halo mass (Taken from Zehavi et al. (2011))**

In terms of the color dependence of galaxy clustering, the trend at fainter luminosities of red galaxies being strongly clustered (with a higher correlation slope, γ, see Fig. 8-8) is due to faint red galaxies being satellite galaxies in relatively massive halos that host bright red central galaxies (Berlind et al. 2005). HOD modeling therefore provides a clear explanation for the increased clustering observed for faint red galaxies. For a given luminosity range ($-20 < M_r < -19$), Zehavi et al. (2011) fit a simplified HOD model with one parameter only to find that the fraction of galaxies that are satellites is much higher for red than for blue galaxies, with ~25% of blue galaxies being satellites and ~60% of red galaxies being satellites. They find that blue galaxies reside in halos with a median mass of $10^{11.7}\ \mathrm{h}^{-1} M_\odot$, while red galaxies reside in higher mass halos with a median mass of $10^{12.2}\ \mathrm{h}^{-1} M_\odot$. However, at a given luminosity, there is not a strong trend between color and halo mass (though there is a strong trend between luminosity and halo mass). Instead, the differences in $w_p(r_p)$ reflect a trend between color and satellite fraction; the increased satellite fraction, in particular, drives the slope of $\xi(r)$ to be steeper for red galaxies compared to blue galaxies. And while the HOD slope α does not change much with increasing luminosity, it does with color, due to the dependence of the satellite fraction on color. Having a higher satellite fraction also places more galaxies in high mass halos (as those host the groups and clusters that contain the satellite galaxies), which increases the large-scale bias and boosts the one-halo term relative to the two-halo term. The HOD model facilitates interpretion of the observed luminosity and color dependence of galaxy clustering and provides strong, crucial constraints on models of how galaxies form and evolve within their parent dark matter halos.

8.4 Interpreting the Evolution of Galaxy Clustering

As mentioned in Sect. 7 above, the galaxies that are observed for clustering measurements at different redshifts are not necessarily the same populations across cosmic time. A significant hurdle in understanding galaxy evolution is knowing how to connect different observed populations at different redshifts. Galaxy clustering measurements can be combined with theoretical models to trace observed populations with redshift in that for a given cosmology, one can model how the clustering of a given population should evolve with time.

The observed evolution of the luminosity dependence of galaxy clustering can be fit surprisingly well using a simple nonparametric, non-HOD, model that relates the galaxy luminosity function to the halo mass function. Conroy et al. (2006) show that directly matching galaxies as a function of luminosity to host halos and subhalos as a function of mass leads to a model for the luminosity-dependent clustering that matches observation from $z \sim 0$ to $z \sim 3$. In this model, the only inputs are the observed galaxy luminosity function at each epoch of interest and the dark matter halo (and subhalo) mass function from N-body simulations. Galaxies are then ranked by luminosity and halos by mass and matched one to one, such that lower luminosity galaxies are associated with halos of lower mass, and galaxies above a given luminosity threshold are assigned to halos above a given mass threshold with the same abundance or number density. This "abundance-matching" method uses as a proxy for halo mass the maximum circular velocity ($V_{\max}$) of the halo; for subhalos, they find that it is necessary to use the value of $V_{\max}$ when the subhalo is first accreted into a larger halo to avoid the effects of tidal stripping. With this simple model, the clustering amplitude and shape as a function of luminosity are matched for SDSS galaxies at $z \sim 0$, DEEP2 galaxies at $z \sim 1$, and Lyman break galaxies at $z \sim 3$. In particular, the clustering amplitude in both the one- and two-halo regimes is well fit, including the deviations from a power law that seen at $z > 1$ (Coil et al. 2006; Ouchi et al. 2005). These results imply a tight correlation between galaxy luminosity and halo mass from $z \sim 0$ to $z \sim 3$.

While abundance-matching techniques provide a simple, zero parameter model for how galaxies populate halos, a richer understanding of the physical properties involved may be gained by performing HOD modeling. Zheng et al. (2007) use HOD modeling to fit the observed luminosity-dependent galaxy clustering at $z \sim 0$ measured in SDSS with that measured at $z \sim 1$, in DEEP2 to confirm that at both epochs there is a tight relationship between the central galaxy luminosity and host halo mass. At $z \sim 1$ the satellite fraction drops for higher luminosities, as at $z \sim 0$, but at a given luminosity, the satellite fraction is higher at $z \sim 0$ than at $z \sim 1$. They also find that at a given central luminosity, halos are ~1.6 times more massive at $z \sim 0$ than $z \sim 1$, and at a given halo, mass galaxies are ~1.4 times more luminous at $z \sim 1$ than $z \sim 0$.

Zheng et al. (2007) further combine these HOD results with theoretical predictions of the growth of dark matter halos from simulations to link $z \sim 1$ central galaxies to their descendants at $z \sim 0$ and find that the growth of both halo mass and stellar mass as a function of redshift depends on halo mass. Lower mass halos grow earlier, which is reflected in the fact that more of their $z \sim 0$ mass is already assembled by $z \sim 1$. A typical $z \sim 0$ halo with mass $3 \times 10^{11}\,\mathrm{h}^{-1} M_{\odot}$ has about 70% of its final mass in place by $z \sim 1$, while a $z \sim 0$ halo with mass $10^{13}\,\mathrm{h}^{-1} M_{\odot}$ has ~50% of its final mass in place at $z \sim 1$. In terms of stellar mass, however, in a $z \sim 0$ halo of mass $5 \times 10^{11}\,\mathrm{h}^{-1} M_{\odot}$, a central galaxy has ~20% of its stellar mass in place at $z \sim 1$, while the fraction rises to ~33% above a halo mass of $2 \times 10^{12}\,\mathrm{h}^{-1} M_{\odot}$. They further find that the mass scale of the maximum star formation efficiency for central galaxies shifts to lower halo mass with time, with a peak of $\sim 10^{12}\,\mathrm{h}^{-1} M_{\odot}$ at $z \sim 1$ and $\sim 6 \times 10^{11}\,\mathrm{h}^{-1} M_{\odot}$ at $z \sim 0$.

At $1 < z < 2$, Wake et al. (2011) use precise photometric redshifts from the NEWFIRM survey to measure the relationship between stellar mass and dark matter halo mass using HOD models. At these higher redshifts, r_0 varies from ~6 to ~11 h^{-1} Mpc for stellar masses $\sim 10^{10}$ to $10^{11}\, M_{\odot}$. The galaxy bias is a function of both redshift and stellar mass and is ~2.5 at $z \sim 1$ and increases to ~3.5 at $z \sim 2$. They find that the typical halo mass of both central and satellite galaxies increases with stellar mass, while the satellite fraction drops at higher stellar mass, qualitatively similar to what is found at lower redshift. They do not find evolution in the relationship between stellar mass and halo mass between $z \sim 2$ and $z \sim 1$ but do find evolution compared to $z \sim 0$. They also find that the peak of star formation efficiency shifts to lower halo mass with time.

Simulations can also be used to connect different observed galaxy populations at different redshifts. An example of the power of this method is shown by Conroy et al. (2008), who compare the clustering and space density of star-forming galaxies at $z \sim 2$ with that of star-forming and quiescent galaxies at $z = 1$ and $z = 0$ to infer both the typical descendants of the $z \sim 2$ star-forming galaxies and constrain the fraction that have merged with other galaxies by $z = 0$. They use halos and subhalos identified in a ΛCDM N-body simulation to determine which halos at $z \sim 2$ likely host star-forming galaxies, and then use the merger histories in the simulation to track these same halos to lower redshift. By comparing these results to observed clustering of star-forming galaxies at $z \sim 1$ and $z \sim 0$, they can identify the galaxy populations at these epochs that are consistent with being descendants of the $z \sim 2$ galaxies. They find that while the lower redshift descendent halos have clustering strengths similar to red galaxies at both $z \sim 1$ and $z \sim 0$, the $z \sim 2$ star forming galaxies cannot all evolve into red galaxies by lower redshift as their space density is too high. There are many more lower redshift descendents than there are red galaxies, even after taking into account mergers. They conclude that most $z \sim 2$ star-forming galaxies evolve into typical L^* galaxies today, while a non-negligible fraction become satellite galaxies in larger galaxy groups and clusters.

In summary, N-body simulations and HOD modeling can be used to interpret the observed evolution of galaxy clustering and further constrain both cosmological parameters and theoretical models of galaxy evolution beyond what can be gleaned from $z \sim 0$ observations alone. They also establish links between distinct observed galaxy populations at different redshifts, allowing one to create a coherent picture of how galaxies evolve over cosmic time.

9 Voids and Filaments

Redshift surveys unveil a rich structure of galaxies as seen in *Fig. 8-3*. In addition to measuring the two-point correlation function to quantify the clustering amplitude as a function of galaxy properties, one can also study higher-order clustering measurements as well as properties of voids and filaments.

9.1 Higher-Order Clustering Measurements

Higher-order clustering statistics reflect both the growth of initial density fluctuations as well as the details of galaxy biasing (Bernardeau et al. 2002), such that measurements of higher-order clustering can test the paradigm of structure formation through gravitational instability as well as constrain the galaxy bias. In the linear regime, there is a degeneracy between the amplitude of fluctuations in the dark matter density field and the galaxy bias in that a highly clustered galaxy population may be biased and trace only the most overdense regions of the dark matter or the dark matter itself may be highly clustered. However, this degeneracy can be broken in the nonlinear regime on small scales. Over time, the density field becomes skewed toward high density as δ becomes greater than unity in overdense regions (where $\delta \equiv (\rho/\bar{\rho}) - 1$) but cannot become negative in underdense regions. Skewness in the galaxy density distribution can also arise from galaxy bias, if galaxies preferentially form in the highest density peaks. One can therefore use the shapes of the galaxy overdensities through measurements of the three-point correlation function to test gravitational collapse versus galaxy bias.

To study higher-order clustering, one needs large samples that cover enormous volumes; all studies to date have focused on low redshift galaxies. Verde et al. (2002) use 2dFGRS to measure the Fourier transform of the three-point correlation function, called the bispectrum, to constrain the galaxy bias without resorting to comparisons with N-body simulations in order to measure the clustering of dark matter. Fry and Gaztanaga (1993) present the galaxy bias in terms of a Taylor expansion of the density contrast, where the first-order term is the linear term, while the second-order term is the nonlinear or quadratic term. Measured on scales of 5–30 h^{-1} Mpc, Verde et al. (2002) find that the linear galaxy bias is consistent with unity ($b_1 = 1.04 \pm 0.11$), while the nonlinear quadratic bias is consistent with zero ($b_2 = -0.05 \pm 0.08$). When combined with the redshift space distortions measured in the two-dimensional two-point correlation function ($\xi(r_p, \pi)$), they measure $\Omega_{\text{matter}} = 0.27 \pm 0.06$ at $z = 0.17$. This constraint on the matter density of the Universe is derived entirely from large-scale structure data alone.

Gaztañaga et al. (2005) measure the three-point correlation function in 2dFGRS for triangles of galaxy configurations with different shapes. Their results are consistent with ΛCDM expectations regarding gravitational instability of initial Gaussian fluctuations. Furthermore, they find that while the linear bias is consistent with unity ($b_1 = 0.93{+}0.10/{-}0.08$), the quadratic bias is nonzero ($b_2/b_1 = -0.34{+}0.11/{-}0.08$). This implies that there is a nongravitational contribution to the three-point function, resulting from galaxy formation physics. These results differ from those of Verde et al. (2002), which may be due to the inclusion by Gaztañaga et al. (2005) of the covariance between measurements on different scales. Gaztañaga et al. (2005) combine their results with the measured two-point correlation function to derive $\sigma_8 = 0.88 + 0.12/-0.10$.

If the density field follows a Gaussian distribution, the higher-order clustering terms can be expressed solely in terms of the lower-order clustering terms. This "hierarchical scaling" holds for the evolution of an initially Gaussian distribution of fluctuations under gravitational instability. Therefore, departures from hierarchical scaling can result either from a non-Gaussian initial density field or from galaxy bias. Redshift space higher-order clustering measurements in 2dFGRS are performed by Baugh et al. (2004) and Croton et al. (2004a), who measure up to the six-point correlation function. They find that hierarchical scaling is obeyed on small scales, though deviations exist on larger scales (~10 h^{-1} Mpc). They show that on large scales, the higher-order terms can be significantly affected by massive rare peaks such as superclusters, which populate the tail of the overdensity distribution. Croton et al. (2004a) also show that the three-point function has a weak luminosity dependence, implying that galaxy bias is not entirely linear. These results are confirmed by Nichol et al. (2006) using galaxies in the SDSS, who also measure a weak luminosity dependence in the three-point function. They find that on scales >10 h^{-1} Mpc, the three-point function is greatly affected by the "Sloan Great Wall," a massive supercluster that is roughly 450 Mpc (Gott et al. 2005) in length and is associated with tens of known Abell clusters. These results show that even 2dFGRS and SDSS are not large enough samples to be unaffected by the most massive, rare structures.

Several studies have examined higher-order correlation functions for galaxies split by color. Gaztañaga et al. (2005) find a strong dependence of the three-point function on color and luminosity on scales <6 h^{-1} Mpc. Croton et al. (2007) measure up to the five-point correlation function in 2dFGRS for both blue and red galaxies and find that red galaxies are more clustered than blue galaxies in all of the N-point functions measured. They also find a luminosity dependence in the hierarchical scaling amplitudes for red galaxies but not for blue galaxies. Taken together, these results explain why the full galaxy population shows only a weak correlation with luminosity.

9.2 Voids

In maps of the large-scale structure of galaxies, voids stand out starkly to the eye. There appear to be vast regions of space with few, if any, L^* galaxies. Voids are among the largest structures observed in the Universe, spanning typically tens of h^{-1} Mpc.

The statistics of voids – their sizes, distribution, and underdensities – are closely tied to cosmological parameters and the physical details of structure formation (e.g., Sheth and van de Weygaert 2004). While the two-point correlation function provides a full description of clustering for a Gaussian distribution, departures from Gaussianity can be tested with higher-order correlation statistics and voids. For example, the abundance of voids can be used to test the non-Gaussianity of primordial perturbations, which constrains models of inflation

(Kamionkowski et al. 2009). Additionally, voids provide an extreme low-density environment in which to study galaxy evolution. As discussed by Peebles (2001), the lack of galaxies in voids should provide a stringent test for galaxy formation models.

9.2.1 Void and Void Galaxy Properties

The first challenge in measuring the properties of voids and void galaxies is defining the physical extent of individual voids and identifying which galaxies are likely to be in voids. The "void finder" algorithm of El-Ad and Piran (1997), which is based on the point distribution of galaxies (i.e., does not perform any smoothing), is widely used. This algorithm does not assume that voids are entirely devoid of galaxies and identifies void galaxies as those with three or less neighboring galaxies within a sphere defined by the mean and standard deviation of the distance to the third nearest neighbor for all galaxies. All other galaxies are termed "wall" galaxies (see ➋ *Fig. 8-13*). An individual void is then identified as the maximal sphere that contains only void galaxies. This algorithm is widely used by both theorists and observers.

Cosmological simulations of structure formation show that the distribution and density of galaxy voids are sensitive to the values of Ω_{matter} and Ω_{Λ} (Kauffmann et al. 1999). Using ΛCDM

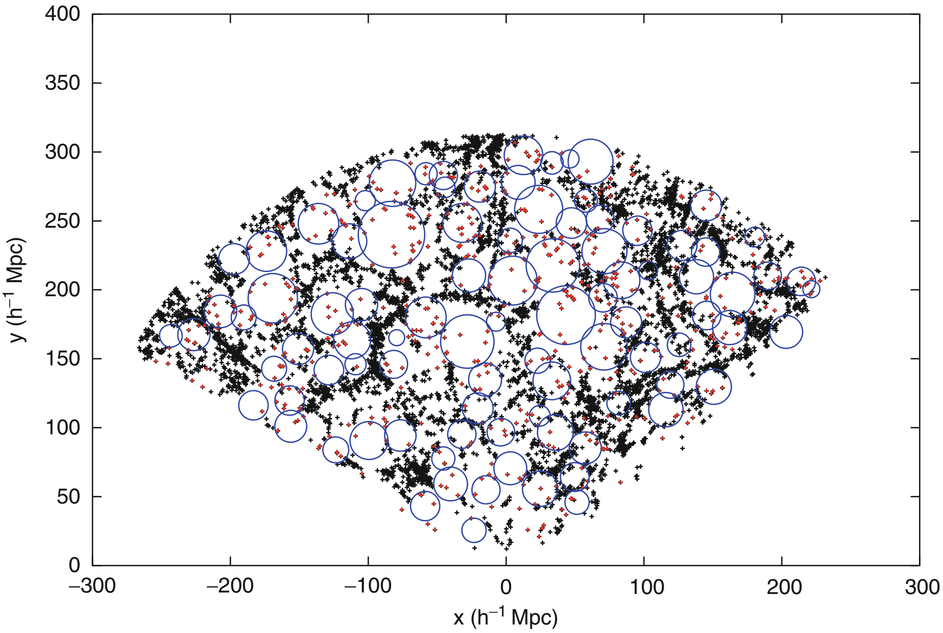


Fig. 8-13

Void and wall galaxies in the SDSS. Shown is a projection of a 10 h^{-1} Mpc slab with wall galaxies plotted as black crosses and void galaxies plotted as red crosses. *Blue circles* indicate the intersection of the maximal sphere of each void with the midplane of the slab (From Pan et al. 2012)

N-body dark matter simulations, Colberg et al. (2005) study the properties of voids within the dark matter distribution and predicts that voids are very underdense (though not empty) up to a well-defined sharp edge in the dark matter density. They predict that 61% of the volume of space should be filled by voids at $z = 0$ compared to 28% at $z = 1$ and 9% at $z = 2$. They also find that the mass function of dark matter halos in voids is steeper than in denser regions of space.

Using similar ΛCDM N-body simulations with a semi-analytic model for galaxy evolution, Benson et al. (2003) show that voids should contain both dark matter and galaxies and that the dark matter halos in voids tend to be low mass and therefore contain fewer galaxies than in higher density regions. In particular, at density contrasts of $\delta < -0.6$, where $\delta \equiv (\rho/\rho_{\mathrm{mean}}) - 1$, both dark matter halos and galaxies in voids should be anti-biased relative to dark matter. However, galaxies are predicted to be more underdense than the dark matter halos, assuming simple physically motivate prescriptions for galaxy evolution. They also predict the statistical size distribution of voids, finding that there should be more voids with smaller radii ($<10\,\mathrm{h}^{-1}$ Mpc) than larger radii.

The advent of the 2dFGRS and SDSS provided the first very large samples of voids and void galaxies that could be used to robustly measure their statistical properties. Applying the "void finder" algorithm on the 2dFGRS dataset, Hoyle and Vogeley (2004) find that the typical radius of voids is $\sim 15\,\mathrm{h}^{-1}$ Mpc. Voids are extremely underdense, with an average density of $\delta\rho/\rho = -0.94$, with even lower densities at the center, where fewer galaxies lie. The volume of space filled by voids is ~40%. Probing an even larger volume of space using the SDSS dataset, Pan et al. (2012) found a similar typical void radius and concluded that ~60% of space is filled by voids, which have $\delta\rho/\rho = -0.85$ at their edges. Voids have sharp density profiles in that they remain extremely underdense to the void radius, where the galaxy density rises steeply. These observational results agree well with the predictions of ΛCDM simulations discussed above.

Studies of the properties of galaxy in voids allow an understanding of how galaxy formation and evolution progresses in the lowest density environments in the Universe, effectively pursuing the other end of the density spectrum from cluster galaxies. Void galaxies are found to be significantly bluer and fainter than wall galaxies (Rojas et al. 2004). The luminosity function of void galaxies shows a lack of bright galaxies but no difference in the measured faint end slope (Croton et al. 2005; Hoyle et al. 2005), indicating that dwarf galaxies are not likely to be more common in voids. The normalization of the luminosity function of wall galaxies is roughly an order of magnitude higher than that of void galaxies; therefore, galaxies do exist in voids just with a much lower space density. Studies of the optical spectra of void galaxies show that they have high star-formation rates, low 4,000Å spectral breaks indicative of young stellar populations, and low stellar masses, resulting in high specific star formation rates (Rojas et al. 2005).

However, red quiescent galaxies do exist in voids, just with a lower space density than blue, star-forming galaxies (Croton et al. 2005). Croton and Farrar (2008) show that the observed luminosity function of void galaxies can be replicated with a ΛCDM N-body simulation and simple semi-analytic prescriptions for galaxy evolution. They explain the existence of red galaxies in voids as residing in the few massive dark matter halos that exist in voids. Their model requires some form of star-formation quenching in massive halos ($>\sim 10^{12}\,M_{\odot}$), but no additional physics that operates only at low density needs to be included in their model to match the data. It is therefore the shift in the halo mass function in voids that leads to different galaxy properties, not a change in the galaxy evolution physics in low-density environments.

9.2.2 Void Probability Function

In addition to identifying individual voids and the galaxies in them, one can study the statistical distribution of voids using the void probability function (VPF). Defined by White (1979), the VPF is the probability that a randomly placed sphere of radius R within a point distribution will not contain any points (i.e., galaxies, see ❯ *Fig. 8-14*). The VPF is defined such that it depends on the space density of points; therefore, one must be careful when comparing datasets and simulation results to ensure that the same number density is used. The VPF traces clustering in the weakly nonlinear regime, not in the highly nonlinear regime of galaxy groups and clusters.

Benson et al. (2003) predict using ΛCDM simulations that the VPF of galaxies should be higher than that of dark matter, that voids as traced by galaxies are much larger than voids traced by dark matter. This results from the bias of galaxies compared to dark matter in voids and the fact that in this model the few dark matter halos that do exist in voids are low mass and therefore often do not contain bright galaxies. Croton et al. (2004b) measure the VPF in the 2dFGRS dataset and find that it follows hierarchical scaling laws in that all higher-order correlation functions can be expressed in terms of the two-point correlation function. They find that even on scales of ~30 h^{-1} Mpc, higher-order correlations have an impact and that the VPF of galaxies is observed to be different than that of dark matter in simulations.

Conroy et al. (2005) measure the VPF in SDSS galaxies at $z \sim 0.1$ and DEEP2 galaxies at $z \sim 1$ and find that voids traced by redder and/or brighter galaxy populations are larger than

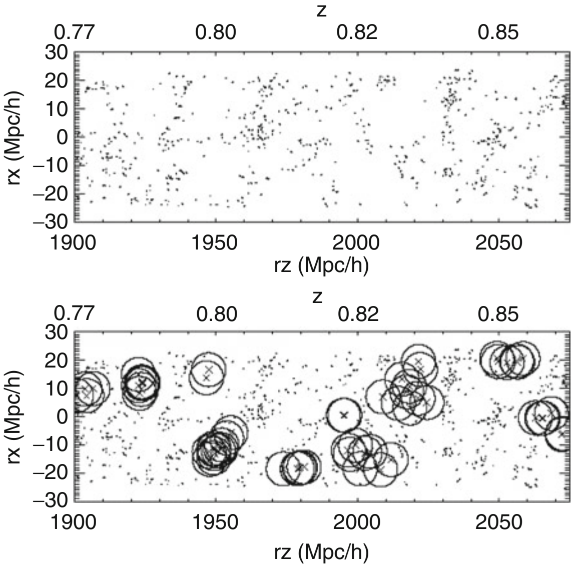


Fig. 8-14

A schematic of the void probability function (VPF). The *top panel* shows the comoving distribution of galaxies in a small portion of the DEEP2 survey (projected through 10 h^{-1} Mpc), while the *lower panel* shows a fraction of the empty spheres identified with a radius of 6 h^{-1} Mpc in the same volume (from Conroy et al. 2005). Because the figure is projected through one dimension, it may appear that galaxies reside inside of identified voids; in three dimensions, the voids contain no galaxies

voids traced by bluer and/or fainter galaxies. They also find that voids are larger in comoving coordinates at $z \sim 0.1$ than at $z \sim 1$; i.e., voids grow over time, as expected. They show that the differences observed in the VPF as traced by different galaxy populations are entirely consistent with differences observed in the two-point correlation function and space density of these galaxy populations. This implies that there does not appear to be additional higher-order information in voids than in the two-point function alone. They also find excellent agreement with predictions from ΛCDM simulations that include semi-analytic models of galaxy evolution.

Tinker et al. (2008) interpret the observed VPF in galaxy surveys in terms of the halo model (see ➲ Sect. 8). They compare the observed VPF in 2dFGRS and SDSS to halo model predictions constrained to match the two-point correlation function and number density of galaxies using a model in which the dark matter halo occupation depends on mass only. They find that with this model, they can match the observed data very well, implying that there is no need for the suppression of galaxy formation in voids, i.e., galaxy formation does not proceed differently in low-density regions. They find that the sizes and emptiness of voids show excellent agreement with predictions of ΛCDM models for galaxies at low redshift to luminosities of $L \sim 0.2L^*$.

9.3 Filaments

Galaxy filaments – long strings of galaxies – are the largest systems seen in maps of large-scale structure and as such, provide a key test of theories of structure formation. Aragon et al. (2010) show that in simulations while filaments occupy only ~10% of the volume of space, they account for ~40% of the mass content of the $z = 0$ Universe. Measuring the typical and maximal length of filaments, as well as their thickness and average density therefore constrains theoretical models. Various statistical methods have been proposed to identify and characterize the morphologies and properties of filaments (e.g., Sousbie et al. 2008 and references therein; see ➲ *Fig. 8-15*).

In terms of their sizes, the largest length scale at which filaments are statistically significant, and hence identified as real objects, is 50–80 h^{-1} Mpc, according to an analysis of galaxies in the Las Campanas Redshift Survey (LCRS; Shectman et al. 1996) by Bharadwaj et al. (2004). They show that while there appear to be filaments in the survey on longer scales, these arise from chance alignments and projection effects and are not real structures. Sousbie et al. (2008) identify and study the length of filaments in SDSS by identifying ridges in the galaxy distribution using the Hessian matrix ($\partial^2\rho/\partial x_i \partial x_j$) and its eigenvalues (see ➲ *Fig. 8-12*). They find excellent agreement between observations and ΛCDM numerical predictions for a flat, low $\Omega_{\rm matter}$ Universe. They argue that filament measurements are not highly sensitive to observational effects such as redshift space distortions, edge effects, incompleteness, or galaxy bias, which makes them a robust test of theoretical models.

Bond et al. (2010) use the eigenvectors of the Hessian matrix of the smoothed galaxy distribution to identify filaments in both SDSS data and ΛCDM simulations and find that the distribution of filaments lengths is roughly exponential, with many more filaments of length $\lesssim 10\,h^{-1}$ Mpc than $>20\,h^{-1}$ Mpc. They find that the filament width distribution agrees between the SDSS data and N-body simulations. The mean filament width depends on the smoothing length; for smoothing scales of 10 and h^{-1} Mpc, the mean filament widths are 5.5 and 8.4 h^{-1} Mpc. In ΛCDM simulations, they find that the filamentary structure in the dark matter density distribution is in place by $z = 3$, tracing a similar pattern of density ridges. This is in contrast to what is found for voids, which become much more prominent and low density at later cosmic epochs.

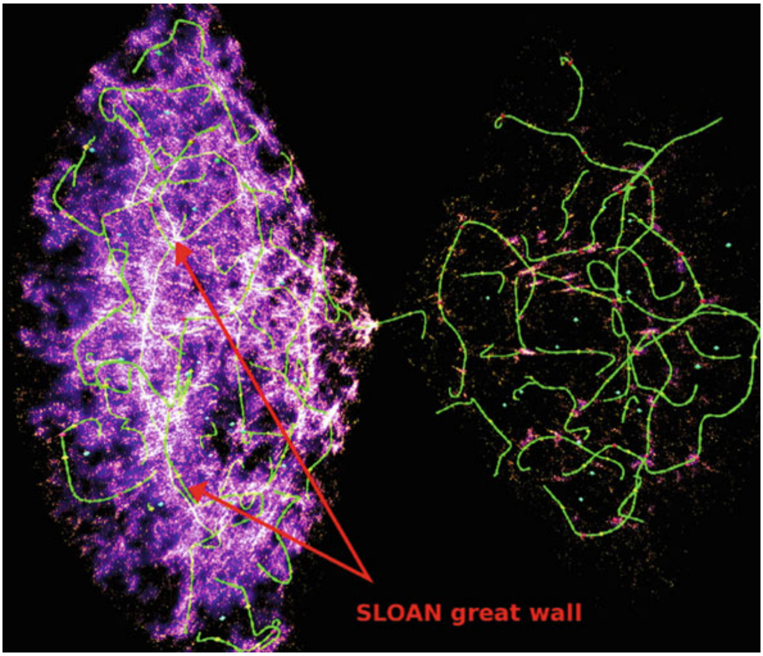


Fig. 8-15
Filaments identified in the SDSS galaxy distribution (from Sousbie et al. 2008). Individual filaments are shown in *green* overlaid on the galaxy density field show in *purple*. The Sloan Great Wall is identified in the foreground, lying between the *red arrows*

Choi et al. (2010) use the methods of Bond et al. (2010) to study the evolution of filamentary structure from $z \sim 0.8$ to $z \sim 0.1$ using galaxies from the DEEP2 survey and the SDSS. They find that neither the space density of filaments nor the distribution of filament lengths has changed significantly over the last 7 Gyr of cosmic time, in agreement with ΛCDM numerical predictions. The distribution of filament widths has changed, however, in that the distribution is broader at lower redshift and has a smaller typical width. This observed evolution in the filament width distribution naturally results from nonlinear growth of structure and is consistent with the results on voids discussed above in that over time, voids grow larger while filaments become tighter (i.e., have a smaller typical width) though not necessarily longer.

10 Summary and Future

This overview of our current understanding of the large-scale structure of the Universe has shown that quantitative measurements of the clustering and spatial distribution of galaxies have wide applications and implications. The nonuniform structure reveals properties of both the galaxies and the dark matter halos that comprise this large-scale structure. Statistics such as the two-point correlation function can be used not only to constrain cosmological parameters but also to understand galaxy formation and evolution processes. The advent of extremely large redshift surveys with samples of hundreds of thousands of galaxies has led to very precise measurements of the clustering of galaxies at $z \sim 0.1$ as a function of various galaxy properties such as luminosity, color, and stellar mass, influencing our understanding of how galaxies form and evolve. Initial studies at higher redshift have revealed that many of the general correlations that are observed between galaxy properties and clustering at $z \sim 0$ were in place when the Universe was a fraction of its current age. As larger redshift surveys are carried out at higher redshifts, much more can be learned about how galaxy populations change with time. Theoretical interpretations of galaxy clustering measurements such as the halo occupation distribution model have also recently made great strides in terms of statistically linking various properties of galaxies with those of their host dark matter halos. Such studies reveal not only how light traces mass on large scales but how baryonic mass and dark matter coevolve with cosmic time.

There are many exciting future directions for studies of galaxy clustering and large-scale structure. Precise cosmological constraints can be obtained using baryon acoustic oscillation signatures observed in clustering measurements from wide-area surveys (Eisenstein et al. 2005). Specific galaxy populations can be understood in greater detail by comparing their clustering properties with those of galaxies in general. For example, the clustering of different types of active galactic nuclei (AGN) can be used to constrain the AGN fueling mechanisms, lifetimes, and host galaxy populations (Coil et al. 2009). As discussed above, measurements of galaxy clustering have the power to place strong constraints on contemporary models of galaxy formation and evolution and advance our understanding of how galaxies populate and evolve within dark matter halos.

Acknowledgments

The author thanks James Aird, Mirko Krumpe, and Stephen Smith for providing comments on earlier drafts of the text.

References

Adelberger, K. L., et al. 2005, ApJ, 619, 697

Aragon, van de Weygaert, & Jones, 2010, MNRAS, 408, 2163

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

Baugh, C. M. 1996, MNRAS, 280, 267

Baugh, C. M., et al. 2004, MNRAS, 351, L44

Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M., & Lacey, C. G. 2000, MNRAS, 311, 793

Benson, A. J., Hoyle, F., Torres, F., & Vogeley, M. S. 2003, MNRAS, 340, 160

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Berlind, A. A., et al. 2005, ApJ, 629, 625

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, PhR, 367, 1

Bharadwaj, S., Bhavsar, S. P., & Sheth, J. V. 2004, ApJ, 606, 25

Bond, N. A., Strauss, M. A., & Cen, R. 2010, MNRAS, 409, 156

Brodwin, M., et al. 2008, ApJL, 687, L65

Choi, E., et al. 2010, MNRAS, 406, 320

Coil, A. L., et al. 2006, ApJ, 644, 671

Coil, A. L., et al. 2008, ApJ, 672, 153

Coil, A. L., et al. 2009, ApJ, 701, 1484

Colberg, J. M., et al. 2005, MNRAS, 360, 216

Colless, M., et al. 2001, MNRAS, 328, 1039

Colless, M., 2004, Cosmological Results from the 2dF Galaxy Redshift Survey, Measuring and Modeling the Universe, ed. Freedman W., Carnegie Observatory Astrophysics Series Vol. 2, Cambridge University Press, 196

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, ApJ, 647, 201

Conroy, C., et al. 2005, ApJ, 635, 990

Conroy, C., et al. 2008, ApJ, 679, 1192

Croton, D. J., & Farrar, G. R. 2008, MNRAS, 386, 2285

Croton, D. J., Norberg, P., Gaztañaga, E., & Baugh, C. M. 2007, MNRAS, 379, 1562

Croton, D. J., et al. 2004a, MNRAS, 352, 1232

Croton, D. J., et al. 2004b, MNRAS, 352, 828

Croton, D. J., et al. 2005, MNRAS, 356, 1155

Davis, M., Geller, M. J., & Huchra, J. 1978, ApJ, 221, 1

Davis, M., Huchra, J., Latham, D. W., & Tonry, J. 1982, ApJ, 253, 423

Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465

Dressler, A. 1980, ApJ, 236, 351

Eisenstein, D. J., et al. 2005, ApJ, 633, 560

El-Ad, H., & Piran, T. 1997, ApJ, 491, 421

Fisher, K. B., Davis, M., Strauss, M. A., Yahil, A., & Huchra, J. P. 1994, MNRAS, 267, 927

Fry, J. N. 1996, ApJL, 461, L65

Fry, J. N., & Gaztanaga, E. 1993, ApJ, 413, 447

Gaztañaga, E., Norberg, P., Baugh, C. M., & Croton, D. J. 2005, MNRAS, 364, 620

Geller, M. J., & Huchra, J. P. 1989, Science, 246, 897

Gott, III, J. R., Dickinson, M., & Melott, A. L. 1986, ApJ, 306, 341

Gott, III, J. R., et al. 2005, ApJ, 624, 463

Gregory & Thompson, 1978, ApJ, 222, 784

Guzzo, L., et al. 2008, Nature, 451, 541

Hamilton, A. J. S. 1992, ApJL, 385, L5

Hamilton, A. J. S. 1993, ApJ, 417, 19

Hamilton, A. J. S. 1998, in ASSL Vol. 231: The Evolving Universe, 185

Hoyle, F., Rojas, R. R., Vogeley, M. S., & Brinkmann, J. 2005, ApJ, 620, 618

Hoyle, F., & Vogeley, M. S. 2004, ApJ, 607, 751

Hubble, E. 1934, ApJ, 79, 8

Hubble, E. P. 1926, ApJ, 64, 321

Joeveer, Einasto, & Tago, 1978, MNRAS, 185, 357

Kaiser, N. 1984, ApJL, 284, L9

Kaiser, N. 1987, MNRAS, 227, 1

Kamionkowski, M., Verde, L., & Jimenez, R. 2009, JCAP, 1, 10

Kauffmann, G., Colberg, J. M., Diaferio, A., & White, S. D. M. 1999, MNRAS, 303, 188

Kerscher, M., Szapudi, I., & Szalay, A. S. 2000, ApJL, 535, L13

Kirshner, R. P., Oemler, A., Jr., & Schechter, P. L. 1978, AJ, 83, 1549

Kravtsov, A. V., et al. 2004, ApJ, 609, 35

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Li, C., et al. 2006, MNRAS, 368, 21

Ma, C.-P., & Fry, J. N. 2000, ApJ, 543, 503

Maddox, S. J., et al. 2010, A&A, 518, L11

Madgwick, D. S., et al. 2003, MNRAS, 344, 847

Mann, R. G., Peacock, J. A., & Heavens, A. F. 1998, MNRAS, 293, 209

Meneux, B., et al. 2006, A&A, 452, 387

Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493

Nichol, R. C., et al. 2006, MNRAS, 368, 1507

Norberg, P., et al. 2001, MNRAS, 328, 64

Ouchi, M., et al. 2004, ApJ, 611, 685

Ouchi, M., et al. 2005, ApJL, 635, L117

Pan, D. C., Vogeley, M. S., Hoyle, F., Choi, Y.-Y., & Park, C. 2012, Cosmic voids in Sloan Digital Sky Survey Data Release 7, MNRAS, 421, 926–934, doi:10.1111/j.1365-2966.2011.20197.x

Peacock, J. A., & Smith, R. E. 2000, MNRAS, 318, 1144

Peacock, J. A., et al. 2001, Nature, 410, 169

Peebles, P. J. E. 1975, ApJ, 196, 647

Peebles, P. J. E. 1980, The Large-Scale Structure of the Universe (Princeton, NJ: Princeton University Press)

Peebles, P. J. E. 2001, ApJ, 557, 495

Postman, M., Lauer, T. R., Szapudi, I., & Oegerle, W. 1998, ApJ, 506, 33

Roche, N., & Eales, S. A. 1999, MNRAS, 307, 703

Rojas, R. R., Vogeley, M. S., Hoyle, F., & Brinkmann, J. 2004, ApJ, 617, 50

Rojas, R. R., Vogeley, M. S., Hoyle, F., & Brinkmann, J. 2005, ApJ, 624, 571

Sandge & Tammann, 1975, ApJ, 196, 313

Sargent, W. L. W., & Turner, E. L. 1977, ApJL, 212, L3

Scoccimarro, R. . 2000, ApJ, 544, 597

Seldner, M., Siebers, B., Groth, E. J., & Peebles, P. J. E. 1977, AJ, 82, 249

Seljak, U. 2000, MNRAS, 318, 203

Shane, C. D., & Wirtanen, C. A. 1967, in Pub. Lick Obs. Vol. 22, part 1

Shapley, H., & Ames, A. 1932, Ann. Harv. Coll. Obs., 88, 41

Shectman, S. A., et al. 1996, ApJ, 470, 172

Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1

Sheth, R. K., & Tormen, G. 1999, MNRAS, 308, 119

Sheth & van de Weygaert, 2004, MNRAS, 350, 517

Sousbie, T., et al. 2008, ApJL, 672, L1

Tegmark, M., & Peebles, P. J. E. 1998, ApJL, 500, L79

Tegmark, M., et al. 2004, ApJ, 606, 702

Tinker, J. L., et al. 2008, ApJ, 686, 53

Verde, L., et al. 2002, MNRAS, 335, 432

Wake, D. A., et al. 2011, ApJ, 728, 46

Weinberg, D. H., Davé, R., Katz, N., & Hernquist, L. 2004, ApJ, 601, 1

White, S. D. M. 1979, MNRAS, 186, 145

York, D. G., et al. 2000, AJ, 120, 1579

Zehavi, I., et al. 2002, ApJ, 571, 172

Zehavi, I., et al. 2004, ApJ, 608, 16

Zehavi, I., et al. 2005, ApJ, 630, 1

Zehavi, I., et al. 2011, ApJ, 736, 59

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760

Zwicky, F., Herzog, E., & Wild, P. 1968, Catalogue of Galaxies and of Clusters of Galaxies (Pasadena: California Institute of Technology, 1961–1968)

9 The Distance Scale of the Universe

Wendy L. Freedman · Barry F. Madore
Carnegie Observatories, Pasadena, CA, USA

1 Introduction ***424***

2 Measurement of Distances ***425***

3 Parallaxes ***425***

4 Rotational Parallax Method ***426***

5 The Role of the Large and Small Magellanic Clouds ***426***

6 RR Lyrae Stars ***427***

7 Red Clump Stars ***430***
7.1 The CMD Method, the Horizontal Branch, and Mira Variables 431

8 Planetary Nebula Luminosity Function (PNLF) ***432***
8.1 Globular Cluster Luminosity Functions (GCLF) 433
8.2 Novae 433
8.3 Type II Supernovae: EPM and SEAM 434
8.4 Cepheid Distance Scale 434
8.4.1 Galactic Cepheids with Trigonometric Parallaxes 435

9 The Distance to the Large Magellanic Cloud Based on Cepheids ***436***
9.1 Tip of the Red Giant Branch (TRGB) Method 437
9.2 Maser Galaxies 439
9.2.1 A Maser Distance to NGC 4258 439
9.2.2 Other Distance Determinations to NGC 4258 440
9.2.3 NGC 4258, UGC 3789, and Their Calibration of H_o 441
9.3 Surface Brightness Fluctuation (SBF) Method 441
9.4 Tully–Fisher Relation 442
9.5 Type Ia Supernovae 442

10 The Extragalactic Distance Scale and the Hubble Constant ***445***

Acknowledgments ***447***

References ***447***

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_9, © Springer Science+Business Media Dordrecht 2013

Abstract: We critically review the methods currently being used to determine extragalactic distances. Within the Milky Way, direct parallaxes and traditional main-sequence fitting for both young open clusters and old globular clusters tie us directly to high-luminosity, variable stars used in extragalactic studies: Cepheids and RR Lyrae stars. These, in turn, inform a calibration of the tip of the red giant branch as well as red clump stars. Apart from the Milky Way, we focus on the Large Magellanic Cloud as an important stopping point at which a large variety of possible distance indicators have been evaluated and differentially tested against each other. Beyond the Local Group many stellar distance indicators fall below current detection and/or resolution limits, and they must be replaced with methods employing higher luminosity (and often much more rare) objects. These methods include the properties of nuclear masers, surface brightness fluctuations, the Tully–Fisher relation, and finally various types of supernovae. Ultimately a calibration of the expansion rate of the universe can provide distances using observed recessional velocities scaled by the Hubble constant.

1 Introduction

Modern extragalactic astronomy began with Edwin Hubble's discovery of Cepheid variables in NGC 6822 (Hubble 1925), M33 (Hubble 1926), and M31 (Hubble 1929). It has taken the better part of a century to develop the instrumentation and techniques to measure distances to accuracies of better than 10%, but this is now routine. Smith (1982) gives a history of the "Great Debate" that occupied the first three decades of the twentieth Century, while Webb (1999) gives a popular and very readable account of measuring the universe. Recent developments and a discussion of the convergence of the extragalactic distance to better than 10% accuracy are reviewed by Freedman and Madore (2010). The current article builds upon and augments this discussion. Previous accounts can be found in Hodge (1982), Rowan-Robinson (1985), Huchra (1992), Jacoby et al. (1992), van den Bergh (1992), Jackson (2007), and Tammann et al. (2008). There are also several conference proceedings dedicated to the extragalactic distance scale, including those edited by van den Bergh and Pritchet (1988) and by Livio et al. (1997); "Stellar Candles for the Extragalactic Distance Scale" edited by Alloin and Gieren (2003) is particularly comprehensive. Finally, a very up-to-date and authoritative account of distance measurements in astronomy has recently been published by de Grijs (2011).

Over the years, a great many extragalactic distance indicators have been tested, calibrated, and applied to nearby galaxies. Until recently these individual determinations have been scattered throughout the literature and were not being tracked or compiled in any systematic way. This has all changed with the introduction of NED-D, which is a feature of the NASA/IPAC Extragalactic Database (NED) that is dedicated to archiving published distances to galaxies, linking them back to the individual objects, and making them widely available in electronic form. These up-to-date compilations can be accessed on object-by-object basis through the main NED interface (http://ned.ipac.caltech.edu/forms/d.html), and the entire archive of distances can be downloaded as a single file from http://ned.ipac.caltech.edu/Library/Distances/. NED-D is updated several times a year.

2 Measurement of Distances

Measuring extragalactic distances generally involves use of one of two types of cosmological distances: the luminosity distance,

$$d_{\mathrm{L}} = \sqrt{\frac{L}{4\pi F}} \tag{9.1}$$

which relates the observed flux (integrated over all frequencies), F, of an object to its intrinsic luminosity, L, emitted in its rest frame; and the angular diameter distance,

$$d_{\mathrm{A}} = \frac{D}{\theta} \tag{9.2}$$

which relates the apparent angular size of an object in radians, θ, to its proper size, D. The luminosity and angular diameter distances are related by

$$d_{\mathrm{L}} = (1+z)^2 d_{\mathrm{A}}. \tag{9.3}$$

The distance modulus, μ, is related to the luminosity distance as follows:

$$\mu \equiv m - M = 5 \log d_{\mathrm{L}} - 5 \tag{9.4}$$

where m and M are the apparent and absolute magnitudes of the objects, respectively, and $\mathrm{d_L}$ is in units of parsecs.

Characterized as "standard candles" and "standard rulers," or more generally known simply as "distance indicators," methods that transcend geometry usually rely on identifying a quantity that is independent of distance (a color, a period, a morphological feature, etc.) that can be precisely measured and then shown to be a predictor of another property of the object in question (say a star or an entire galaxy) that is either dimmed in luminosity or reduced in size with distance.

3 Parallaxes

Measuring a parallax (i.e., using direct geometric triangulation with the annual displacement of the Earth around the Sun as a baseline) is the most straightforward way of determining interstellar distances. In principle, triangulation can be used over any distance; however, the level of precision required in its application is quickly outstripped by the increasing distances encountered across our Milky Way galaxy and beyond. Virtually all of the distances discussed from this point on build upon and use trigonometric parallaxes as foundational but generally rely on the inverse-square law fall-off of apparent luminosity with distance as the means of inferring distances across cosmic scales. A comprehensive review of trigonometric parallaxes and especially the impact of the Hipparcos astrometric satellite can be found in Heck and Caputo (1999), amplified and updated in Perryman (2009). We discuss the HST Cepheid parallax calibration in ➲ Sect. 8.4.1.

4 Rotational Parallax Method

In a variety of contexts (optical and radio), the combination of proper motions and radial velocities has been implemented as a novel and promising means of determining a geometric distance to nearby galaxies (Loeb et al. 2005). For objects in the Local Group (at distances out to ~1 Mpc.), proper motions having a precision of 10–20 μarcsec/year measured from the ground (in the radio) or from space (e.g., with the planned European Space Agency mission GAIA) are sufficient to provide "rotational parallaxes" within a 3–5-year time baseline. For a rotating disk galaxy (such as M31 or M33), the in-plane rotational velocity (km/s) can be equated to the transverse angular velocity (i.e., the proper motion) appropriately scaled by the distance. By measuring (tangential) proper motions and having knowledge of the radial-velocity field of the disk (corrected for inclination), one can derive the distance assuming reasonable symmetries in the structure of the disk and its velocity field.

Using 22 GHz H_2O (water) masers, a rotational parallax for M33 has already been measured (Brunthaler et al. 2005), and a complementary program is now underway using recently discovered water masers in M31 (Darling 2011). Water masers have also been observed in IC 10 (Henkel et al. 1986), but given their less than regular velocity field (Wilcots and Miller 1998), the inversion of proper motions and radial velocities to determine a distance to this galaxy will be more challenging. Perhaps the apparent angular divergence of sources (discussed by Darling 2011 for M31) as IC 10 approaches the Milky Way at −350 km/s, can be used in the future.

5 The Role of the Large and Small Magellanic Clouds

The Large Magellanic Cloud has played and probably will continue to play a central role in the refinement of the extragalactic distance scale. The primary reason is, of course, its proximity. Additionally, the LMC is a composite system, containing a mix of old, intermediate-age, and young stars in a system that is sufficiently massive that it has numerous examples of almost every type of distance indicator currently in use; one notable exception being a recent type Ia supernova. An extensive compilation of 275 distance estimates to the LMC is currently available from the NASA/IPAC Extragalactic Database (NED): http://nedwww.ipac.caltech.edu/cgi-bin/nDistance?name=lmc. The range in quoted distances is almost certainly dominated by systematic errors. The values for the LMC distance modulus fall generally in the range of 18.1–18.7 mag (i.e., 42–55 kpc), with more recent values tending to cluster around a distance modulus of 18.5 mag (see Alves 2004; Schaeffer 2008).

The LMC is sufficiently far away that its stellar components can all be considered to be at the same distance. Under this assumption, most of the observed dispersion in any luminosity-based distance indicator can be ascribed either to differential reddening or an intrinsic scatter in the distance determination method itself. Similarly, differences in relative zero points of various distance indicators when compared in a single system, like the LMC, can expose systematic errors in one or both of the methods. At a finer level of scrutiny, however, the back-to-front geometry of the LMC is in fact measurable by individual distance indicators of exceptionally high internal precision, such as the Classical Cepheids (Gascoigne and Shobbrook 1978), or by methods employing large statistical samples (e.g., Weinberg and Nikolaev 2001). Systematic shifts in the apparent distance modulus with position on the sky can then, with some certainty, be attributed to tilt of the disk of the galaxy with respect to the plane of the sky.

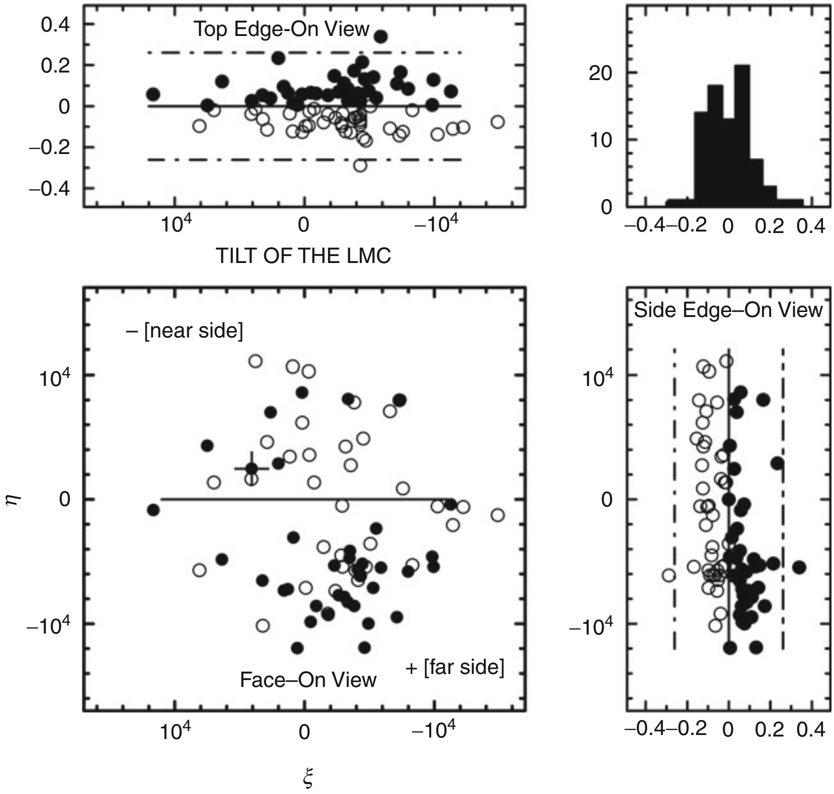


Fig. 9-1
The residuals in the Leavitt law at 3.6 μm recently observed using Spitzer plotted as a function of position in the LMC (Scowcroft et al. 2012) from which the tilt of the LMC can be measured

Figure 9-1 illustrates this for Cepheids in the LMC recently observed using mid-IR data Cepheid from Spitzer (Scowcroft et al. 2012).

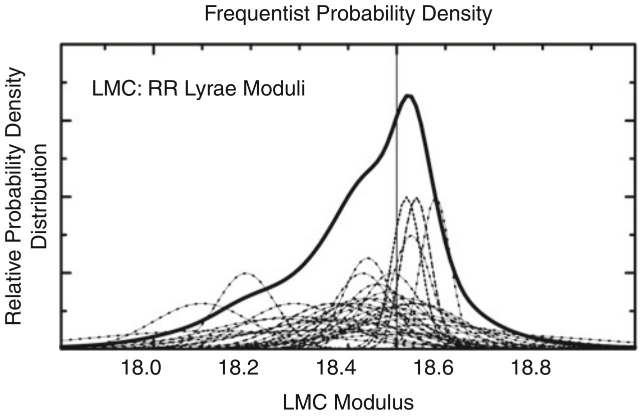
The Small Magellanic Cloud is only slightly (0.4 mag or 10 kpc) further away than the LMC, and it is also somewhat less massive, but its three-dimensional geometry is so distorted that the system as a whole holds less promise than the LMC (or even some more distant galaxies like IC 1613, M31, and M33) in testing distance indicators. The SMC appear to be so tidally disrupted that the line-of-sight differences in distance moduli across the system appears to be as large as 0.5 mag peak-to-peak.

6 RR Lyrae Stars

It is probable that every type of variable star has been tested for its suitability as a distance indicator, if for no other reason than its period (if it has a stable period) or its characteristic timescale might be used as a distance-independent means of predicting the star's intrinsic luminosity. RR Lyrae stars are no exception to this rule, but they also have the added advantage that many of them are known to be members of globular clusters. This latter attribute offers up the possibility that they can be independently calibrated in absolute terms, through main sequence fitting, for example. On the negative side, for most extragalactic applications, RR Lyrae variables are faint. They reside on the horizontal branch and have absolute magnitudes that are around $M_V = +0.5$ mag. An up-to-date and comprehensive review of the properties of RR Lyrae stars can be found in the monograph by Smith (1995), while a recent review of their status as distance indicators can be found in Bono (2003). An indication of the systematics encountered in applying RR Lyraes to the distance scale can be seen in ➋ *Fig. 9-2*, which compares 38 independent RR Lyrae distance determinations to the LMC.

Heroic ground-based efforts to detect RR Lyrae stars beyond the Magellanic Cloud were surpassed only in precision and sample size by HST. However, no effort by any telescope on the ground or in space has been made to detect RR Lyraes beyond the Local Group; they are simply too faint. Nevertheless 23 Local Group galaxies have been successfully surveyed for RR Lyrae stars; a representative sampling of these is given in ➋ *Table 9-1*.

If RR Lyraes are to be continued to be observed in Local Group galaxies and used to test for consistency in overlapping distance indicators, then there are some newly revealed advantages in moving the calibration to the near infrared. As modeled by Catelan et al. (2004) and earlier observed by Longmore et al. (1986), the period-luminosity relation for RR Lyrae stars becomes significantly better defined in the near infrared, having a steeper slope and much decreased intrinsic scatter, as compared to the optical where the V-band magnitude is almost degenerate with period. This behavior for variable stars in general is discussed in Madore and Freedman (2012).



◘ Fig. 9-2
Thirty-eight RR Lyrae star distance determinations to the LMC. The modal value is 18.52 mag. *Dashed lines* are unit-area Gaussians whose mean is at the published distance and whose sigma corresponds to the published statistical error. Determinations without a published uncertainty were all assigned an uncertainty of 0.10 mag for plotting purposes only. The *solid line* represents the frequentist sum of these Gaussians. The *thin solid vertical line* marks a fiducial distance modulus of 18.50 mag acting as a point of reference when comparing subsequent plots

Table 9-1
RR Lyrae distances

Galaxy	(m-M) ± err	NED reference code
M31	23.23 ± 0.15	1988ApJ...331..135P
	23.35 ± 0.15	1987ApJ...316..517P
	23.46 ± 0.11	2009AJ....138..184S
	23.48 ± 0.11	2004AJ....127.2738B
	23.49 ± 0.19	2010ApJ...708..817F
	23.50 ± 0.10	2004AJ....127.2738B
	23.51 ± 0.11	2004AJ....127.2738B
	23.52 ± 0.08	2009ApJ...704L.103C
	23.59 ± 0.19	2010ApJ...708..817F
	23.64 ± 0.19	1988ApJ...331..135P
M33	24.67 ± 0.08	2006AJ....132.1361S
	24.47 ± 0.12	1992MmSAI..63..331L
	24.84 ± 0.16	2000AJ....120.2437S
NGC 147	23.85 ± 0.22	1987AJ.....94.1556S
	23.92 ± 0.25	1990AJ....100..108S
	24.16 ± 0.16	2010ApJ...708..293Y
NGC 205	24.65 ± 0.25	1992AJ....103...84S
NGC 6822	23.36 ± 0.17	2003ApJ...588L..85C
IC 10	24.56 ± 0.08	2008ApJ...688L..69S
IC 1613	24.10 ± 0.27	1992AJ....104.1072S
	24.30 ± 0.05	2003AJ....125.1261D
	24.32 ± 0.16	2001ApJ...550..554D
	24.44 ± 0.10	2010ApJ...712.1259B
	24.47 ± 0.12	2010ApJ...712.1259B
LMC	18.19 ± 0.06	1990AJ....100.1532W
	18.52 ± 0.02	2000AAp...363L...1K
	18.48 ± 0.08	2004AAp...423...97B
	18.61 ± 0.28	1999AAp...348L..33G
SMC	18.86 ± 0.07	1988AJ.....96..872W
	18.78 ± 0.15	1986MNRAS.221..887R
	18.86 ± 0.07	2003AJ....125.1261D
	18.93 ± 0.24	2004AJ....128..736W
And I	24.49 ± 0.06	2005AJ....129.2232P
And II	24.11 ± 0.02	2004AJ....127..318P
And III	24.38 ± 0.06	2005AJ....129.2232P
And VI	24.56 ± 0.06	2002AJ....124.1464P
Draco	19.40 ± 0.02	2004AJ....127..861B
Fornax	20.53 ± 0.09	2007ApJ...670..332G
	20.66 ± 0.03	2003MNRAS.345..747M
Leo A	24.52 ± 0.09	2003AJ....125.1261D
Sextans	24.52 ± 0.09	2003AJ....125.1261D
Sculptor	19.67 ± 0.02	2008AJ....135.1993P

Table 9-2
Galactic Cepheids with geometric parallaxes

Cepheid	P(days)	logP	μ (mag)	σ (%)	Distance (pc)
RT Aur	3.728	0.572	8.15	7.9	427
T Vul	4.435	0.647	8.73	12.1	557
FF Aql	4.471	0.650	7.79	6.4	361
δ Cep	5.366	0.730	7.19	4.0	274
Y Sgr	5.773	0.761	8.51	13.6	504
X Sgr	7.013	0.846	7.64	6.0	337
W Sgr	7.595	0.881	8.31	8.8	459
β Dor	9.842	0.993	7.50	5.1	316
ζ Gem	10.151	1.007	7.81	6.5	365
l Car	35.551	1.551	8.56	9.9	515

7 Red Clump Stars

After leaving the hydrogen-burning main sequence, stars can enter a second fairly long-lived phase of central nuclear energy generation, this time powered by core helium burning. The helium-burning main sequence of low-mass stars is seen as the horizontal branch typical of many Population II systems. The helium-burning main sequence of intermediate-age (higher-mass) clump stars can be seen in the color–magnitude diagram as an enhancement in the luminosity function near to the giant branch and at an absolute magnitude level slightly brighter than the older, traditional horizontal branch.

Use of the red giant clump as a distance indicator has had a checkered history. This is primarily due to the difficulty of establishing the age and metallicity of the underlying stellar populations, but it is also affected by the changing wavelength domain that has been used to calibrate it. The evolving nature of our understanding of this feature is best illustrated by its application to determining a distance to the LMC.

As with many distance indicators, age and metallicity effects are often in play at the same time, but the luminosity measured at different wavelengths may be responding, to varying degrees, to each of these effects. The question ultimately boils down to this: Is the luminosity of any given feature in the color–magnitude diagram a strong function of age and/or metallicity? Often a second parameter is usually sought to "calibrate out" the sensitivity of the luminosity to nuisance parameters such as metallicity and age (see, e.g., Alves and Sarajedini 1999; Girardi and Salaris 2001). However, it is also true that the degree to which any one of these parameters affects the luminosity can itself be a function of wavelength. Atmospheric line transitions that are often sensitive to and diagnostic of metallicity are heavily concentrated in the shorter-wavelength blue and optical portions of the spectrum; therefore by moving to the red or near infared, one should become naturally less sensitive to metallicity and thus a better distance indicator. The other immediate benefit of moving as far to the infrared as possible is the monotonically decreasing effect of interstellar reddening.

Udalski et al. (1998) and later Udalski (2000) used I-band data for the red clump stars in the LMC to argue for a very small distance modulus to the LMC of 18.24 mag, or nearly 0.25 mag short of the generally accepted distance modulus of 18.50 mag. The debate rapidly gathered momentum and diverged on several fronts. Stanek et al. (1998) compared I-band

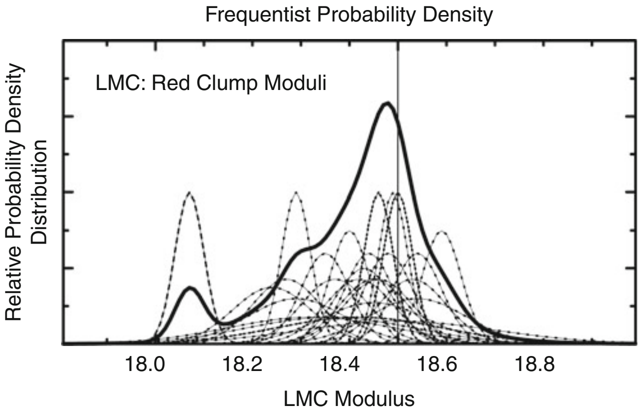


Fig. 9-3
Thirty-two Red Clump distance determinations to the LMC. The modal value is 18.52 mag (See RR Lyrae caption (*Fig. 9-2*) for plotting details)

data of red clump stars from the Hipparcos catalog with their counterparts in two fields in the LMC and found an even shorter distance to the LMC corresponding to a true modulus of only 18.07 mag ± 0.03 (random) ± 0.09 (systematic).

Alves (2000) then moved the debate to the near infrared by providing a K-band calibration based again on galactic stars with direct individual parallaxes from Hipparcos combined with I-band photometry. Koerwer (2009) subsequently applied a mid-IR calibration of the red clump to the LMC and found 18.54 ± 0.06 mag. Alves et al. (2002) and also Pietrzynski et al. (2003) each determined a distance modulus for the LMC that yielded a value of 18.50 ± 0.03 mag. The Alves calibration was claimed to be independent of metallicity over the range $-0.5 < [Fe/H] < 0.00$ dex. However, Girardi and Salaris (2001) have modeled corrections for population effects, involving age and metallicity, (see also Sarajedini 1999 and Pietrzynski et al. 2010 for complementary empirical approaches) and find corrections in the I band of 0.2 mag. For completeness, see also Seidel et al. (1987), Paczynski and Stanek (1998), Grocholski and Sarajedini (2002) and Sarajedini et al. (2002).

Sixteen distance determinations to the LMC using the red clump method have been published; their values range from 18.03 to 18.59 mag; there distribution is shown in *Fig. 9-3*. Eleven other galaxies, all in the Local Group, have also had their distances estimated by the red clump method.

7.1 The CMD Method, the Horizontal Branch, and Mira Variables

Given full color–magnitude diagrams (CMD) for the resolved populations of nearby galaxies not only can individual stars (e.g., Cepheids, RR Lyraes, Miras, etc.) and featured groups of stars (such as the TRGB and the horizontal branch) be used to gauge distances, but the entire CMD itself can be deconstructed, with distances and stellar populations being self-consistently solved for. The concept finds its origins with Tolstoy and Saha (1996) and has been more fully developed by Dolphin (2002) and applied most recently by de Jong et al. (2008). Only eight

Local Group galaxies (And I, II, and III, Sculptor, CVn II, LGS 3, Ursa Minor, and Fornax) are close enough to use the horizontal branch method to determine distances. And four galaxies (LMC, M33, NGC 5128, and the Phoenix Dwarf) have been monitored sufficiently that they have had distance estimates derived from long-period Mira variables.

8 Planetary Nebula Luminosity Function (PNLF)

The distinctive and rapid cutoff observed at the bright end of the [OIII]5007λ emission-line luminosity function for planetary nebulae provides a distance determination method (e.g., Ciardullo et al. 1989). Examples of some luminosity functions obtained at the current distance limits of this method's application and the adopted fits are given in Fig. 9-4. Extensively

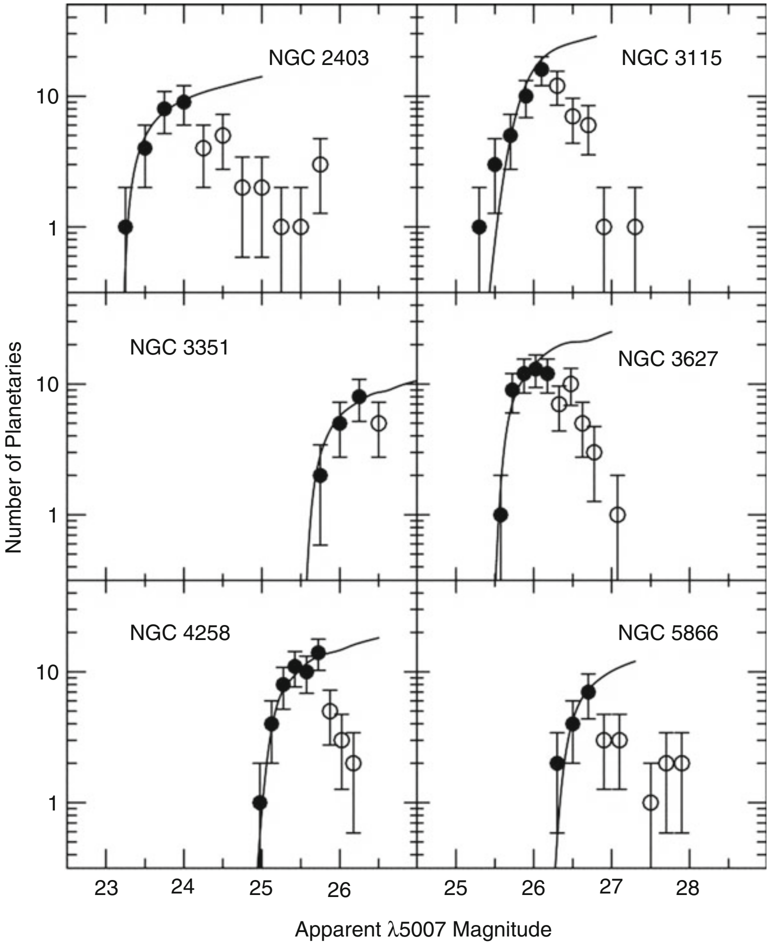


Fig. 9-4
Examples of distance determination fits to planetary nebula luminosity functions (From Ciardullo et al. 2002)

referenced reviews of the technique can be found in Jacoby et al. (1992), Jacoby and Ciardullo (1993), and more recently Ciardullo (2003).

One important advantage of the PLNF method is that it can be applied to both elliptical galaxies and Population II rich spiral galaxies, thereby providing an important point of direct comparison of distance determination methods that are otherwise heavily restricted to only one or the other populations, such as the SBF method for elliptical galaxies and the Cepheid PL relation for spirals. An excellent reviews of PLNF distances are given in Ciardullo et al. (2002) and Ciardullo (2005).

There are some subtleties that need to be paid attention to in using the PLNF method for deriving distances. These include culling out interlopers, especially at the bright end of the luminosity function where the measurement of the edge is intrinsically sensitive to small-number statistics. Known contaminants include HII regions, supernova remnants, background high-redshift, emission-line galaxies, and rare but super-luminous planetary nebulae. Accounting for the possible effects of interstellar extinction is a concern not easily dealt with (Ciardullo et al. 2002).

To date, 57 galaxies have had their PNLF measured and used to estimate their distances; the most distant application being made to objects in the Virgo and Fornax clusters. Pressing the technique to large distances requires not only larger aperture telescopes but also custom-built (or tuneable) narrowband (30-Å) filters targeting specific redshift intervals appropriate to individual galaxies.

8.1 Globular Cluster Luminosity Functions (GCLF)

At one time, it was suggested that the brightest globular cluster in any given system might be used as a distance indicator (Sandage 1968). This proved not to be the case as later studies found that the there was no bright cutoff in the globular cluster luminosity function (GCLF) and that the brightest globular clusters simply scaled with the size of the total population. Later the magnitude of the peak of the luminosity function, which is relatively bright ($M_V \sim -7.5$ mag), was explored as a distance indicator (Hanes 1979). Recent reviews and discussions of the GCLF method can be found in Tamman and Sandage (1999), Ferrarese et al. (2000), and Richler (2003). Further complications include the realization that cluster destruction rates will be different in disk galaxies compared to elliptical galaxies and will be a function of an individual cluster's physical structure, such as compactness. Gnedin and Ostriker (1998) calculate that 50–90% of a globular cluster population will be destroyed, by dynamical friction, tidal shocks, or simple core collapse and disintegration, in a Hubble time. Furthermore, the discovery of color bimodality in the integrated color–magnitude distributions of globular clusters in a number of nearby galaxies (see Forbes et al. 1997) means that there will be age or metallicity-induced differences in the mean magnitude of the luminosity function itself, thereby complicating its calibration and compromising its application as a distance indicator (see Richler 2003 for an in-depth discussion). To date, over 142 galaxies have their distances estimated by the GCLF method.

8.2 Novae

Novae are cataclysmic variables (thought to result from thermonuclear runaway on a mass-accreting white dwarf in a close binary system) whose absolute visual magnitudes cover a wide range but can reach up to $M_V = -10$ mag. This alone makes them of interest as potential distance indicators. Zwicky (1936) was the first to note that the peak brightness of a nova correlates with its rate of decline; that is, intrinsically faint novae fade more slowly than their brighter counterparts. As most recently emphasized by Ferrarese et al. (2003), a calibration of the maximum magnitude versus rate of decline (MMRD) relation (actually called the "life-luminosity" relation by Zwicky) has proven to be "remarkably illusive," primarily due to a lack of data. Other than M31, where the heroic surveys by Arp (1956) early on, and then again later by Capaccioli et al. (1989) produced many novae, the LMC (with only a handful of well-studied novae, Capaccioli et al. 1990) is the only other galaxy that has a viable sample of published light curves with which to attempt a calibration (e.g., Della and Livio 1995 for a recent example).

Novae cannot be predicted, but once found, they fade rapidly and are not easily followed up at other wavelengths or at higher signal-to-noise. Moreover, the interpretation of single-band discoveries is subject to being systematically compromised by unknown amounts of individual lines-of-sight extinction. A 24-orbit HST campaign to discover and follow novae in the Virgo-cluster galaxy, Messier 49 = NGC 4472 (Ferrarese et al. 2003), resulted in only five "fairly complete" light curves. However, the sobering conclusion of that study was that there are substantial differences in the shape of the MMRD in M31, the Milky Way, and M49. A more optimistic assessment of novae as extragalactic distance indicators can be found in Gilmozzi and Della Valle (2003) but see also Della and Livio (1995). Only three galaxies (LMC, M31, and M100 = NGC 4321) have had their distances estimated using novae.

8.3 Type II Supernovae: EPM and SEAM

Two methods are used to determine distances in the universe based on type II SN: the expanding photosphere method (EPM, Kirschner and Kwan 1974; Eastman and Kirshner 1989) and the spectral-fitting expanding atmosphere method (SEAM, Baron et al. 1995, 2004). The EPM is based on the Baade–Wesselink method (Baade 1926) and is particularly effective when the metallic line-blanketing effects in the optical are small (early phases). EPM uses a black-body approximation to the spectral energy distribution, adjust by a variety of correction factors (Eastman et al. 1996). SEAM uses synthetic spectra and fits the observed energy distributions directly. For a discussion of the relative merits of the two techniques, the reader is referred to the recent paper on the SN II distance scale (Dessart and Hiller 2005). Over 30 galaxies have had their distances determined using type II supernovae.

8.4 Cepheid Distance Scale

Since the discovery of the Leavitt Law (Leavitt 1908; Leavitt and Pickering 1912) and its use by Hubble to measure the distances to the Local Group galaxies, Cepheid variables have remained a widely applicable and powerful method for measuring distances to nearby galaxies. Cepheids' periods of pulsation range from 2 to over 100 days, and their intrinsic brightnesses range from $-2 < \mathrm{M}_V < -6$ mag. Detailed reviews of the Cepheid distance scale and its calibration can be found in Madore and Freedman (1991), Sandage and Tammann (2006), Fouque et al. (2007), and Barnes (2009). A recent, more lengthy discussion is contained in Freedman and Madore (2010), while a review of the early history of the subject is given in Fernie (1969).

There are many steps that must be taken in applying Cepheids to the extragalactic distance scale. Overcoming crowding and confusion is the key to the successful discovery, measurement, and use of Cepheids in galaxies beyond the Local Group. From the ground, atmospheric turbulence degrades the image resolution, decreasing the contrast of point sources against the background. As higher precision data have been accumulated for Cepheids in greater numbers and in different physical environments, it has become possible to search for and investigate a variety of lower level, but increasingly important, systematics affecting the Leavitt Law.

The physical basis for the Leavitt Law is well understood. Cepheid pulsation occurs because of the changing atmospheric opacity with temperature in the doubly ionized helium zone. This zone acts like a heat engine and valve mechanism. During the portion of the cycle when the ionization layer is opaque to radiation that layer traps energy resulting in an increase in its internal pressure. This added pressure acts to elevate the layers of gas above it, resulting in the observed radial expansion. As the star expands, it does work against gravity and the gas cools. As it does so, its temperature falls back to a point where the doubly ionized helium layer recombines and becomes transparent again, thereby allowing more radiation to pass. Without that added source of heating, the local pressure drops, the expansion stops, the star recollapses, and the cycle repeats. The alternate trapping and releasing of energy in the helium ionization layer ultimately gives rise to the periodic change in radius, temperature, and luminosity seen at the surface. Not all stars are unstable to this mechanism. The cool (red) edge of the Cepheid instability strip is thought to be controlled by the onset of convection, which then prevents the helium ionization zone from driving the pulsation. For hotter temperatures, the helium ionization zone is located too far out in the atmosphere for significant pulsations to occur. Further details can be found in the classic stellar pulsation textbook by Cox (1980), and Freedman and Madore (2010).

Cepheids have been intensively modeled numerically, with increasingly sophisticated hydrodynamical codes (for a recent review, see Buchler 2009). While continuing progress is being made, the challenges remain formidable in following a dynamical atmosphere and in modeling convection with a time-dependent mixing length approximation. In general, observational and theoretical period–luminosity–color relations are in reasonable agreement (e.g., Caputo 2008). However, subtle effects (e.g., that of metallicity on the luminosities and colors of Cepheids) remain difficult to predict from first principles.

8.4.1 Galactic Cepheids with Trigonometric Parallaxes

An accurate trigonometric parallax calibration for galactic cepheids has been long sought but very difficult to achieve in practice. All known classical (galactic) Cepheids are more than 250 pc away; therefore for direct distance estimates good to 10%, parallax accuracies of ±0.2 milliarcsec are required, necessitating space observations. The HIPPARCOS satellite reported parallaxes for 200 of the nearest Cepheids, but (with the exception of Polaris) even the best of these had very low signal-to-noise ratio (Feast and Catchpole 1997).

Benedict et al. (2007) used the Fine Guidance Sensors (FGS) on HST to provide the first high-precision, geometric parallaxes to ten nearby galactic Cepheids having periods ranging from 4 to 36 days. *Spitzer* mid-infrared data for the HST parallax calibration sample, as well as Cepheids in the LMC, have now been obtained (Monson et al. 2012; Scowcroft et al. 2012; Freedman et al. 2012). The advantages of the mid-infrared are many, including the small dispersion in the mid-infrared Leavitt relations, as well as the insensitivity to reddening and metallicity.

9 The Distance to the Large Magellanic Cloud Based on Cepheids

Several thousand Cepheids have been identified and cataloged in the LMC (Leavitt 1908; Alcock et al. 2000; Soszynski et al. 2008), all at essentially the same distance. Both historically and today, the slope of the Leavitt Law is both statistically and systematically better determined in the LMC than it is for Cepheids in our own galaxy. This is especially true for the long-period end of the calibration where the extragalactic samples are much larger than the small sample of nearby Cepheids in the Milky Way. The main drawback to using the LMC as the fundamental calibrator of the Leavitt Law is the fact that the LMC Cepheids are of lower metallicity than many of the more distant spiral galaxies useful for measuring the Hubble constant. This systematic is largely eliminated by adopting the higher-metallicity galactic calibration or calibration based on the distance to the maser spiral galaxy, NGC 4258.

In *Fig. 9-5*, we show the Leavitt Law at 3.6 μm for 82 Cepheids in the LMC (with 6 < P < 60 days) as given by Scowcroft et al. (2012). The dispersion in the 3.6-μm relation amounts to only ±0.106 mag (or an uncertainty of ±5% in distance for a single Cepheid). For comparison, we also show the V-band data from Madore and Freedman (1991) for LMC Cepheids. The dispersion in this case is more than a factor of 2 greater, amounting to ±0.252 mag. One hundred previously published distance moduli to the LMC, based solely on Cepheids, are shown in *Fig. 9-6*.

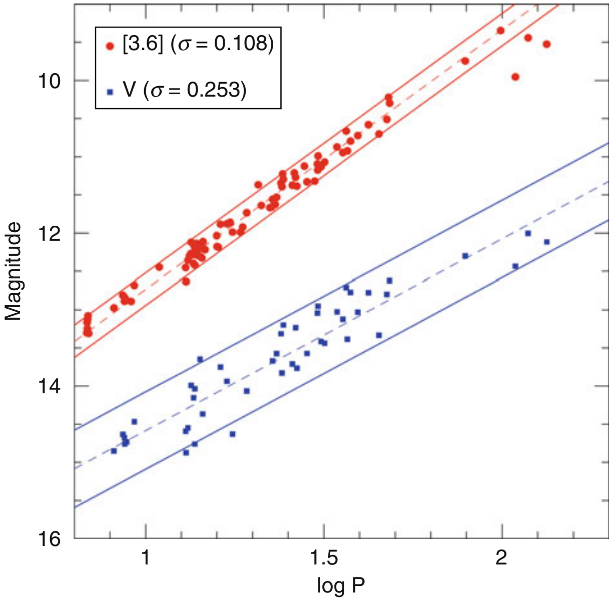


Fig. 9-5

Phase-averaged 3.6-μm (*red circles*) and V-band (*blue squares*) Leavitt Law relations for the LMC. The 3.6-μm data are from Scowcroft et al. (2012), the V-band from Madore and Freedman (1991). Note the small dispersion of ±0.11 mag at 3.6 μm, which is more than a factor of 2 less than for the V-band. The *dashed lines* represent weighted least square fits to the PL relations for Cepheids in the period range 6–60 days. The *solid lines* denote 2-σ ridge lines

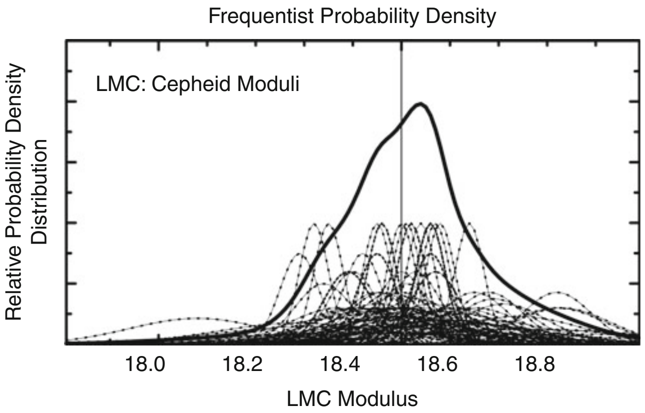


Fig. 9-6
One hundred Cepheid distance moduli to the LMC. The modal value of the Cepheid distribution function falls at about 18.55 mag

9.1 Tip of the Red Giant Branch (TRGB) Method

The tip of the red giant branch (TRGB) method uses the theoretically well-understood and observationally well-defined discontinuity in the luminosity function of stars evolving up the red giant branch in old, metal-poor stellar populations. This feature has been calibrated using galactic globular clusters, and because of its simplicity and straightforward application, it has been widely used to determine distances to nearby galaxies. The method was developed quantitatively in two papers: one by DaCosta and Armandroff (1990) for galactic globular clusters and the other by Lee et al. (1993), where the use of a quantitative digital filter to measure the tip location was first introduced in an extragalactic context. The method has a precision comparable to Cepheids. Recent and excellent reviews of the topic have been published by Rizzi et al. (2007) and Bellazzini (2008).

Approximately 250 galaxies have had their distances measured by the TRGB method. This is to be compared to a total of 57 galaxies with Cepheid distances. (A comprehensive compilation of direct distance determinations is available at the following website: http://nedwww.ipac.caltech.edu/level5/NED1D/ned1d.html). A comparison of nine applications of the TRGB method to the LMC is given in ➋ *Fig. 9-7*. In practice, the TRGB method is observationally a much more efficient technique since, unlike for Cepheid variables, there is no need to follow them through a variable light cycle; a single-epoch observation, made at two wavelengths (to provide color information), is sufficient. A recent example of applying the TRGB technique to the maser galaxy, NGC 4258, is shown in ➋ *Figs. 9-7* and ➋ *9-8*.

The TRGB is also well-understood theoretically (e.g., Iben and Renzini 1983; Freedman and Madore 2010). In brief, the helium core at the center of a red giant is supported by electron degeneracy pressure. A hydrogen-burning shell surrounds the core and provides the luminosity of the star. Fall-out, in the form of "helium ash" from the shell, increases the mass of the core over time. As the core mass increases, the radius shrinks, the temperature of the shell, and consequently the luminosity generated in the shell, increases; the star rises along the giant branch

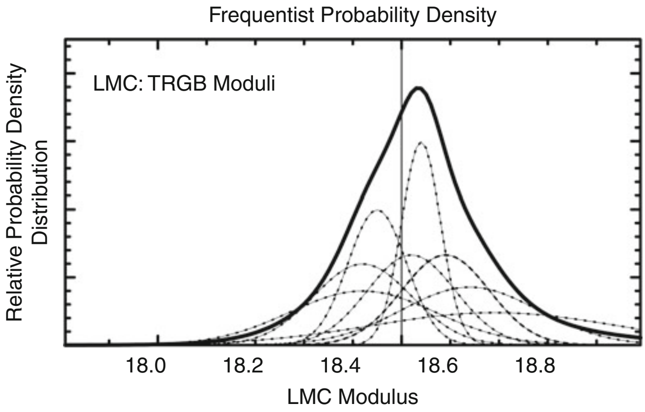


Fig. 9-7

Nine tip of the red giant branch (TRGB) distance determinations to the LMC. The modal value is 18.53 mag. See RR Lyrae caption (*Fig. 9-2*) for plotting details

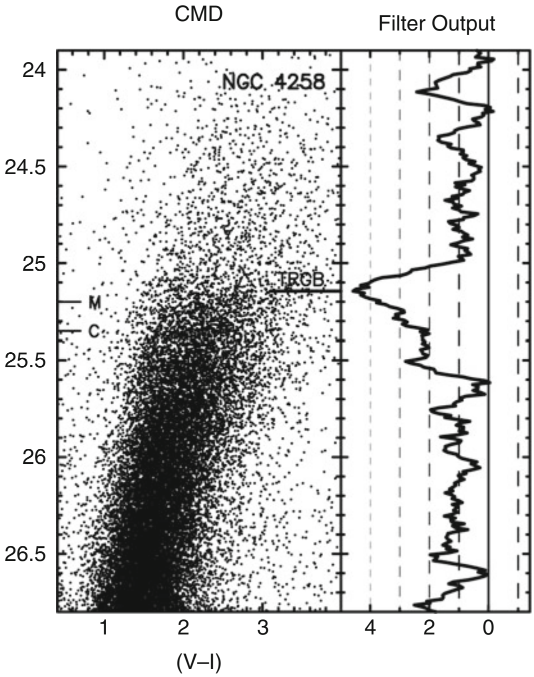


Fig. 9-8

An example of the detection and measurement of the discontinuity in the observed luminosity function for red giant branch stars in the halo of the maser galaxy NGC 4258 (Mager et al. 2008). The color–magnitude diagram on the left has been adjusted for metallicity such that the TRGB is found at the same apparent magnitude independent of color/metallicity of the stars at the tip. The *right panel* shows the output of an edge-detection (modified Sobel) filter whose peak response indicates the TRGB magnitude and whose width is used as a measure of the random error on the detection

with increasing luminosity and higher core temperatures. When the (isothermal) core temperature reaches a physically well-defined temperature, helium ignites throughout the core. This helium core ignition lifts the electron degeneracy within the core. This dramatic change in the equation of state is such that the core flash is internally quenched in a matter of seconds. The core is inflated and settles down to a lower-luminosity, helium core-burning main sequence. The transition from the red giant to the horizontal branch occurs rapidly (within a few million years) so that observationally the TRGB can be treated as a physical discontinuity. Nuclear physics fundamentally controls the stellar luminosity at which the RGB is truncated, essentially independent of the chemical composition and/or residual mass of the envelope sitting above the core.

The radiation from stars at the TRGB is redistributed with wavelength as a function of the metallicity and mass of the envelope. Empirically it is found that the bolometric corrections are smallest in the I-band, and most recent measurements have been made at this wavelength. The small residual metallicity effect on the TRGB luminosity is well documented and can be empirically calibrated out (see Madore et al. 2009).

9.2 Maser Galaxies

H_2O (water) megamasers have been shown to provide an independent and potentially powerful means of accurately measuring extragalactic distances geometrically. Lo (2005) has reviewed both the physical nature of megamasers and their application to the extragalactic distance scale. The technique utilizes the mapping of 22.2 GHz water maser sources orbiting in the accretion disks of black holes in spiral galaxies with active galactic nuclei, where modeling of those disks assumes simple Keplerian motion. A rotation curve is derived for the major axis of the disk; proper motions are measured on the near side of the disk on the minor axis. Scaling the angular velocities across the line of sight to the absolute (radial) velocities along the line of sight yields the distance.

The method requires a sample of accretion disks that are relatively edge on (so that a rotation curve can be obtained from radial-velocity measurements) and a heating source such as x-rays or shocks to produce maser emission. The basic assumption is that the maser emission arises from trace amounts of water vapor ($<10^{-5}$ in number density) in very small density enhancements in the accretion disk and that they act as perfect dynamical test particles. The maser sources appear as discrete peaks in the spectrum or as unresolved spots in the images constructed from Very Long Baseline Interferometry (VLBI). Measurements of the acceleration ($a = V^2/r$) are obtained directly by monitoring the change of maser radial velocities over time from single-dish observations. Proper motions are obtained from observed changes in angular position in interferometer images. The approximately Keplerian rotation curve for the disk can then be modeled, allowing for warps and radial structures. The best studied galaxy, NGC 4258, at a distance of about 7 Mpc is still too close to provide a secure measurement of the Hubble constant on its own (i.e., free from local velocity-field perturbations), but it can serve as an invaluable independent check of the Cepheid zero-point calibration.

9.2.1 A Maser Distance to NGC 4258

VLBI observations of H_2O maser sources surrounding the active galactic nucleus of NGC 4258 reveal them to be in a very thin, differentially rotating, slightly warped disk. The Keplerian velocity curve has deviations of less than 1%. The disk has a rotational velocity in excess of 1,000 km/s at distances on the order of 0.1 pc from the inferred super-massive (10^7 M$_\odot$) nuclear black hole. Detailed analyses of the structure of the accretion disk as traced by the masers have been published (e.g., Herrnstein et al. 1999; Humphreys et al. 2008, and references therein). Over time, it has been possible to measure both proper motions and accelerations of these sources and thereby allow for the derivation of two independent distance estimates to this galaxy. The excellent agreement of these two estimates supports the a priori adoption of the Keplerian disk model and gives distances of 7.2 ± 0.2 and 7.1 ± 0.2 Mpc, respectively.

Because of the simplicity of the structure of the maser system in NGC 4258 and its relative strength, NGC 4258 will remain a primary test bed for studying systematic effects that may influence distance estimates. Several problems may limit the ultimate accuracy of this technique, however. For example, because the masers are only distributed over a small angular part of the accretion disk, it is difficult to assess the importance of noncircular orbits. Of possible concern, eccentric disks of stars have been observed in a number galactic nuclei where the potential is dominated by the black hole, as is the case for NGC 4258. In addition, even if the disk is circular, it is not a given that the masers along the minor axis are at the same radii as the masers along the major axis. The self gravity of the disk also may need to be investigated and modeled since the maser distribution suggests the existence of spiral arms (Humphreys et al. 2008). Finally, radiative transfer effects may cause nonphysical motions in the maser images. Although the current agreement of distances using several techniques is comforting, having only one sole calibrating galaxy for this technique remains a concern, and further galaxies will be required to ascertain the limiting uncertainty in this method.

9.2.2 Other Distance Determinations to NGC 4258

The first Cepheid distance to NGC 4258 was published by Maoz et al. (1999) who found a distance of 8.1 ± 0.4 Mpc, scaled to an LMC-calibrated distance modulus of 18.50 mag. Newman et al. (2001) found a distance modulus of 29.47 ± 0.09 (random) ±0.15 (systematic), giving a distance of 7.83 ± 0.3 ±0.5 Mpc. Macri et al. (2006) reobserved NGC 4258 discovering 281 Cepheids at BV and I wavelengths in two radially (and chemically) distinct fields. Their analysis gives a distance modulus of 29.38 ± 0.04 ± 0.05 mag (7.52 ± 0.16 Mpc), if one adopts $\mu(LMC) = 18.50$ mag. Several more recent determinations of resolved-star (Cepheid and TRGB) distance moduli to NGC 4258 are in remarkably good agreement with the maser distance modulus. For instance, diBenedetto (2008) measures a Cepheid distance modulus of 29.28 ± 0.03 ± 0.03 for NGC 4258 (corresponding to a distance of 7.18 Mpc); Benedict et al. (2007) find a distance modulus of 29.28 ± 08 mag; and Mager et al. (2008) also find a value of 29.28 ± 0.04 ± 0.12 mag both from Cepheids and from the TRGB method. These latter studies are in exact agreement with the current maser distance. Higher accuracy has come from larger samples with higher signal-to-noise data and improved treatment of metallicity.

An alternative approach to utilizing the maser galaxy in the distance scale is to adopt the geometric distance to NGC 4258 as foundational, use it to calibrate the Leavitt Law, and from there, determine the distance to the LMC. Macri et al. (2006) adopted this approach and concluded that the true distance modulus to the LMC is 18.41 ± 0.10 mag.

9.2.3 NGC 4258, UGC 3789, and Their Calibration of H_o

The distance to NGC 4258 has been used to bypass the LMC in calibrating the Cepheid PL relation and then secondary methods. For example, Macri et al. (2006) and Riess et al. (2009a, b) have adopted the distance to NGC 4258 to calibrate the supernova distance scale, as discussed further in ➋ Sect. 10.

Attempts to measure distances to additional megamaser host galaxies have been challenging. About 2000 galaxies have been surveyed in search of nuclear disk masers, with about 100 masers being culled from this sample. The rather low detection rate of 5% is likely due to detection sensitivity, combined with the geometric constraint that the maser disk be viewed nearly edge on, because the maser emission is expected to be highly beamed in the plane of the disk. About 30 of these masers have spectral profiles indicative of emission from thin disks: that is, masers at the galactic systemic velocity and groups of masers symmetrically spaced in velocity. In the end, about a dozen maser galaxies are sufficiently strong that they are candidates for being imaged with phase-referenced VLBI techniques. However, only about five have been found to have sufficiently simple structures that they can be fit to dynamical models and thereby have their distances determined. The most promising example is UGC 03789 at a recessional velocity of 3,325 km/s. A first-epoch determination of a geometric distance to this galaxy has been published by Braatz et al. (2010) as part of the *Megamaser Cosmology Project* (Reid et al. 2009). For its geometric/maser distance of 49.9 ± 7.0 Mpc, this galaxy alone gives $H_o = 69 \pm 11$ km/s/Mpc. Correcting the observed velocity for large-scale flow perturbations due to Virgo, the Great Attractor and the Shapley Concentration give V = 3,530 ± 26 km/s and $H_o = 71$ km/s/Mpc.

If a significant number of maser galaxies can be found and precisely observed even further into the Hubble flow, this method can, in principle, compete with methods such as SNe Ia for directly measuring distances at cosmologically significant scales.

9.3 Surface Brightness Fluctuation (SBF) Method

For distances to elliptical galaxies and early-type spirals with large bulge populations, the surface brightness fluctuation (SBF) method, first introduced by Tonry and Schneider (1988), overlaps with and substantially exceeds the current reach of the TRGB method (Tonry et al. 2001). Both methods use properties of the red giant branch luminosity function to estimate distances. The SBF method quantifies the effect of distance on an overall measure of resolution of the Population II red giant stars, naturally weighted both by their intrinsic luminosities and relative numbers. What is measured is the pixel-to-pixel variance in the photon statistics (scaled by the surface brightness) as derived from an image of a pure population of red giant branch stars. For fixed surface brightness, the variance in a pixel (of fixed angular size) is a function of distance simply because the total number of discrete sources contributing to any given pixel increases with the square of the distance. While the TRGB method relies entirely on the very brightest red giant stars, the SBF method uses a luminosity-weighted integral over the entire RGB population in order to define a typical "fluctuation star" whose mean magnitude, $\overline{M_I}$, is assumed to be universal and can therefore be used to derive distances. For recent discussions of the SBF method, the reader is referred to Biscardi et al. (2008), Blakeslee et al. (2009), and an update by Blakeslee et al. (2010).

Aside from the removal of obvious sources of contamination such as foreground stars, dust patches, and globular clusters, the SBF method does require some additional corrections.

It is well known that the slope of the red giant branch in the color–magnitude diagram is a function of metallicity, and so the magnitude of the fluctuation star is both expected and empirically found to be a function metallicity. A (fairly steep) correction for metallicity has been derived and can be applied using the mean color of the underlying stellar population $\overline{M_I} = -1.74 + 4.5(V - I)_o - 1.15$ (Tonry et al. 2002).

9.4 Tully–Fisher Relation

The total luminosity of a spiral galaxy (corrected to face-on inclination to account for extinction) is strongly correlated with the galaxy's maximum (corrected to edge-on inclination) rotation velocity. This relation, calibrated via the Leavitt Law or TRGB, becomes a powerful means of determining extragalactic distances (Tully and Fisher 1977; Aaronson et al. 1986; Pierce and Tully 1988; Giovanelli et al. 1997). The Tully–Fisher relation at present is one of the most widely applied methods for distance measurements, providing distances to thousands of galaxies both in the general field and in groups and clusters. The scatter in this relation is wavelength-dependent and approximately ±0.3–0.4 mag or 15–20% in distance (Giovanelli et al. 1997; Sakai et al. 2000; Tully and Pierce 2000).

In a general sense, the Tully–Fisher relation can be understood in terms of the virial relation applied to rotationally supported disk galaxies, under the assumption of a constant mass-to-light ratio (Aaronson et al. 1979). However, a detailed self-consistent physical picture that reproduces the Tully–Fisher relation (e.g., Steinmetz and Navarro 1999) and the role of dark matter in producing almost universal spiral galaxy rotation curves (McGaugh et al. 2000) still remain a challenge.

Spitzer archival data have recently yielded an unexpected and exciting discovery. Of the 23 nearby galaxies with HST Cepheid distances that can be used to independently calibrate the Tully–Fisher relation, there are 17 that currently also have 3.6-μm total magnitudes (Seibert et al. 2012). In ➋ *Fig. 9-9* (left three panels), we show the B-, V-, and I-band TF relations for the entire sample of currently available calibrating galaxies from Sakai et al. (2000). Their magnitudes have been corrected for inclination-induced extinction effects, and their line widths have been corrected to edge-on. The scatter is ±0.43, 0.36, and 0.36 mag for the B-, V-, and I-band relations, respectively; the outer lines follow the mean regression at ±2 sigma. In the right panel of ➋ *Fig. 9-7*, we show the mid-IR TF relation for the same 17 galaxies with Cepheid distances and IRAC observations, measured here at 3.6 μm. The scatter at 3.66 μm is ±0.31 mag. This calibration will be applied to Spitzer 3.6-μm data for several hundred galaxies (Seibert et al. 2012).

9.5 Type Ia Supernovae

One of the most accurate means of measuring cosmological distances out into the Hubble flow utilizes the peak brightness of type Ia supernovae (SNe Ia). The potential of supernovae for measuring distances was clear to early researchers (e.g., Baade, Minkowski, Zwicky), but it was the Hubble diagram of Kowal (1968) that set the modern course for this field, followed by decades of work by Sandage, Tammann, and collaborators (e.g., Sandage and Tammann 1982; Sandage and Tammann 1990); see also the review by Branch (1998). Analysis by Pskovskii (1984), followed by Phillips (1993), established a correlation between the magnitude of an SN Ia

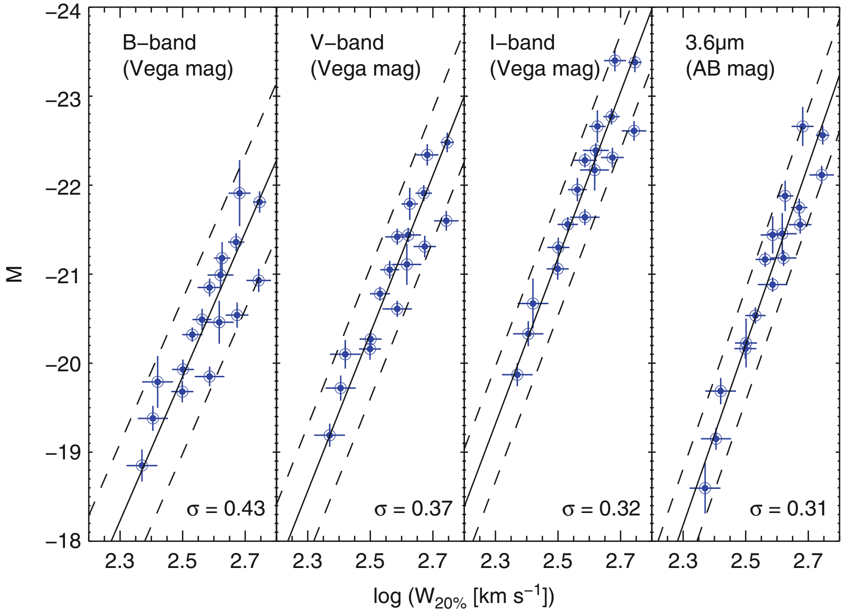
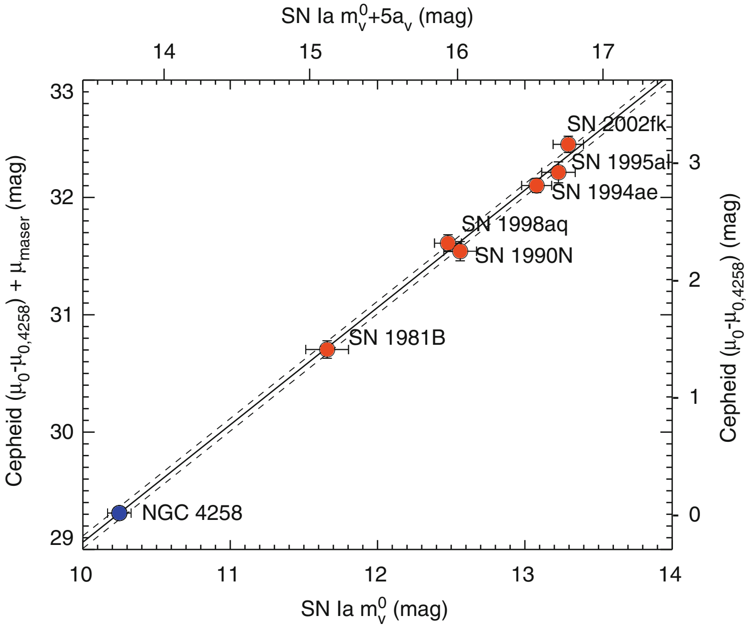


Fig. 9-9
Multiwavelength Tully–Fisher relations. The three *left panels* show the B-, V-, and I-band TF relations for galaxies calibrated with independently measured Cepheid moduli at the end of the HST Key Project. W is the inclination-corrected line width (from Sakai et al. 2000) measured at the 20% power point. The *right-hand panel* shows the TF relation for the subset of galaxies drawn from the Key Project calibrators with measured 3.6-μm total AB magnitudes from Seibert et al. (2012). Data for 17 galaxies are available at all four wavelengths. The dispersions in these relations are shown in the *lower right*

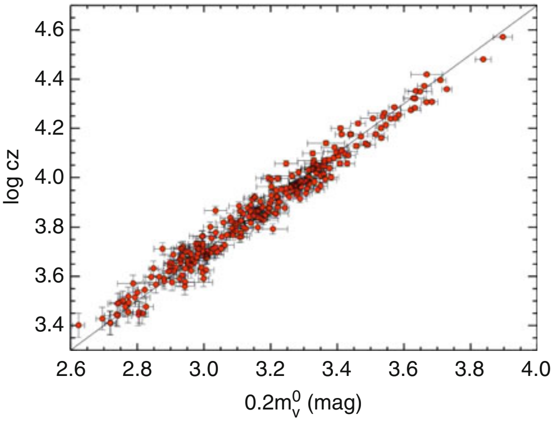
at peak brightness and the rate at which it declines, thus allowing supernova luminosities to be "standardized." This method currently probes farthest into the unperturbed Hubble flow, and it possesses very low intrinsic scatter; in recent studies, the decline-rate corrected SN Ia Hubble diagram is found to have a dispersion of ±7–10% in distance (e.g., Folatelli et al. 2010; Hicken et al. 2009). A simple lack of Cepheid calibrators prevented the accurate calibration of type Ia supernovae for determination of $H_\circ$ prior to HST. Substantial improvements to the supernova distance scale have resulted from recent dedicated, ground-based supernova search and follow-up programs yielding CCD light curves for nearby supernovae (e.g., Hamuy et al. 1996; Jha et al. 2006; Contreras et al. 2010). Sandage and collaborators undertook a major program with HST to find Cepheids in nearby galaxies that have been host to Type Ia supernovae (Sandage et al. 1996; Saha et al. 1999), and thereby provided the first Cepheid zero-point calibration, which has recently been followed up by Macri et al. (2006) and Riess et al. (2009a, b) ➋ *Figs. 9-10* and ➋ *9-11*.

SNe Ia result from the thermonuclear runaway explosions of stars. From observations alone, the presence of SNe Ia in elliptical galaxies suggests that they do not come from massive stars. Many details of the explosion are not yet well understood, but the generally accepted view is that of a carbon–oxygen, electron-degenerate, nearly Chandrasekhar mass white dwarf orbiting in



■ **Fig. 9-10**

A comparison of Cepheid and SNe Ia distances (*red points*), as described in Riess et al. (2009b). The calibrating galaxy, NGC 4258, is added in *blue*



■ **Fig. 9-11**

Supernova Hubble diagram based on 240 supernovae with $z < 0.1$. The sample is from Hicken et al. (2009) and has been used by Riess et al. (2009b) for their determination of H_o

a binary system with a close companion (Whelan and Iben 1973). As material from the Roche lobe of the companion is deposited onto the white dwarf, the pressure and temperature of the core of the white dwarf increase until explosive burning of carbon and oxygen is triggered. An alternative model is that of a "double degenerate" system (merger with another white dwarf).

Although on observational grounds, there appear to be too few white dwarf pairs, this issue has not been conclusively resolved. A review of the physical nature of SNe Ia can be found in Hillebrandt and Niemeyer (2000).

A defining characteristic of observed SNe Ia is the lack of hydrogen and helium in their spectra. It is presumed that the orbiting companion is transferring hydrogen- and helium-rich material onto the white dwarf; however, despite extensive searches, this hydrogen or helium has never been detected, and it remains a mystery as to how such mass transfer could take place with no visible signature. It is not yet established whether this is a problem of observational detection or whether these elements are lost from the system before the explosion occurs.

Various models for SN Ia explosions have been investigated. The most favored model is one in which a subsonic deflagration flame is ignited, which subsequently results in a supersonic detonation wave (a delayed detonation). The actual mechanism that triggers an SN Ia explosion is not well understood: successfully initiating a detonation in a CO white dwarf remains extremely challenging. In recent years, modeling in 3D has begun, given indications from spectropolarimetry that the explosions are not spherically symmetric. The radiative transport calculations for exploding white dwarf stars are complex. However, there is general consensus that the observed (exponential shape of the) light curves of SN e Ia are powered by the radioactive decay of ^{56}Co to ^{56}Fe. The range of observed supernova peak brightnesses appears to be due to a range in ^{56}Ni produced. However, the origin of the peak magnitude – decline rate – is still not well understood.

Despite the lack of a solid theoretical understanding of SNe Ia, empirically they remain one of the best-tested, lowest-dispersion, and highest-precision means of measuring relative distances out into the smooth Hubble flow.

10 The Extragalactic Distance Scale and the Hubble Constant

We now give a brief discussion of the application of the extragalactic distance scale to measurements of the Hubble constant and its uncertainties. A recent, detailed review of the Hubble constant can be found in Freedman and Madore (2010). We focus here on recent efforts, subsequent to that of the Hubble Space Telescope Key Project (Freedman et al. 2001).

A recent calibration of SNe Ia has come from Riess et al. (2009a, b, 2011) from a new calibration of six Cepheid distances to nearby well-observed supernovae using the Advanced Camera for Surveys (ACS) and the Near-Infrared Camera and Multi-Object Spectrometer (NICMOS) on HST. By extending to the near-infrared, these observations of the newly discovered Cepheids directly address the systematic effects of metallicity and reddening. Riess et al. determine a value of $H_o = 74.2 \pm 3.6$ km s^{-1} Mpc^{-1} combining the systematic and statistical errors. This value is in excellent agreement with that from the Key Project (Freedman et al. 2001), which is calibrated using the galactic Cepheid parallax sample. At the current time, there is not much need for larger, low-redshift samples since the dominant remaining uncertainties are systematic rather than statistical. Recent studies (e.g., Wood-Vasey et al. 2008; Folatelli et al. 2010) confirm that supernovae are better standard candles at near-infrared (JHK) wavelengths and minimize the uncertainties due to reddening.

Tammann et al. (2008) also undertook a recent recalibration of supernovae, as well as a comparison of the Cepheid, RR Lyrae, and TRGB distance scales. In contrast, they find a value of $H_{\circ} = 62.3 \pm 4.0\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, where the quoted (systematic) error includes their estimated uncertainties in both the Cepheid and TRGB calibration zero points. Their quoted error is dominated by the systematic uncertainties in the Cepheid zero point and the small number of supernova calibrators, both of which are estimated by them to be at the 3–4% level; however, the H_o values differ by more than 2-σ. A discussion of the reason for the differences in these analyses can be found in Riess et al. (2009a, b); these include the use of more heavily reddened galactic Cepheids, the use of less accurate photographic data, and a calibration involving multiple telescope/instruments for supernovae by Tammann, Sandage, and Reindl.

A recent and comprehensive review of the application of the SBF method to determining cosmic distances, and its comparison to the fundamental plane (FP) method is given in Blakeslee et al. (2002). This analysis leads to the a value of $H_o = 72 \pm 4\,(random) \pm 11\,(systematic)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Mould and Sakai (2008) have used the TRGB as an alternate calibration to the Cepheid distance scale for the determination of H_o. They use 14 galaxies for which TRGB distances can be measured to calibrate the Tully–Fisher relation, and determine a value of $H_{\circ} = 73 \pm 5$ (statistical only) km s^{-1} Mpc^{-1}, a value about 10% higher than found earlier by Sakai et al. (2000) based on a Cepheid calibration of 23 spiral galaxies with Tully–Fisher measurements. In subsequent papers, they calibrate the SBF method (Mould and Sakai 2009a) and the FP for early-type galaxies and the luminosity scale of type Ia supernovae (Mould and Sakai 2009b). They conclude that the TRGB and Cepheid distance scales are all consistent using SBF, FP, SNe Ia, and the TF relation.

As part of a new Carnegie Hubble Project (CHP), new mid-infrared observations of Cepheids have been obtained at 3.6 μm using the *Spitzer Space Telescope* (Freedman et al. 2012). A mid-IR zero point of the Leavitt Law is obtained using time-averaged 3.6-μm data for the five longest-period, high-metallicity (Milky Way) Cepheids having trigonometric parallaxes (Monson et al. 2012). The slope is measured from new time-averaged, mid-IR data for 82 Large Magellanic Cloud Cepheids falling in the period range $0.8 < \log(P) < 1.8$ (Scowcroft et al. 2011). These data yield a value of $\mathrm{H}_{\circ} = 74.3 \pm 1.5$ (statistical) ±2.1 (systematic) km/s/Mpc (Freedman et al. 2012). This *Spitzer* calibration decreases the systematic uncertainty in $\mathrm{H}_{\circ}$ over that obtained by the Hubble Space Telescope Key Project by a factor of 3.

Mid-infrared observations retire many of the systematic uncertainties in the Cepheid distance scale, which dominate at optical wavelengths. Specifically, the mid-IR reduces, by a factor of at least 20, the sensitivity of the Cepheid distance scale to reddening corrections and assumptions about the universality of the reddening law. The new *Spitzer* calibration eliminates instrumental zero-point uncertainties for the Cepheids, as it is based solely on observations taken using a single, stable detector/telescope combination (the *Spitzer* IRAC camera). This calibration is also less sensitive to metallicity effects since it is centered on a high-metallicity (galactic) zero point. Moreover, the 3.6-μm band is demonstrably less sensitive to the atmospheric metallicity effects seen at shorter wavelengths. The current systematic uncertainty on the Hubble constant is now dominated by the small number of galactic calibrators having independent, trigonometric distances. This systematic error will be significantly reduced with the inclusion of addition Cepheid parallaxes expected to be forthcoming from the Global Astrometric Interferometer for Astrophysics (GAIA) mission. In principle, a value of $H_{\circ}$ having a well-tested and robust systematic error budget of only 2% is within reach over the next decade.

Acknowledgments

Many dedicated individuals have contributed to the progress in accurately calibrating the extragalactic distance scale; it has been a community effort spanning the better part of a century. The authors made extensive use of the *NASA/IPAC Extragalactic Database* (NED) in preparing this chapter.

References

Aaronson, M., Huchra, J. P., & Mould, J. R. 1979, ApJ, 229, 1

Aaronson, M., Bothun, G., Mould, J., Huchra, J., Schommer, R. A., & Cornell, M. E. 1986, ApJ, 302, 536

Albrecht, A., Bernstein, G., Cahn, R., Freedman, W. L., Hewitt, J., et al. 2006, astro-ph/0609591

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., et al. 2000, AJ, 119, 2194

Alloin, D., & Gieren, W. 2003, in Stellar Candles for the Extragalactic Distance Scale, Lecture Notes in Physics, ed. D. Alloin, & W. Gieren (Heidelberg: Springer), 635, 1

Alves, R. D. 2000, ApJ, 539, 732

Alves, R. D. 2004, New AR, 48, 659

Alves, R. D., & Sarajedini, A. 1999, ApJ, 511, 225

Alves, R. D., Rejkuba, M., Minniti, D., & Cook, K. H. 2002, ApJL, 573, 51

Arp, H. C. 1956, AJ, 61, 15

Baade, W. 1926, AN, 228, 359

Barnes, T. G. 2009, in Stellar Pulsation: Challenges for Theory and Observation, AIP Conference Proceedings, Vol. 1170, ed. J. A. Guzik, & P. A. Bradley (Melville: American Institute of Physics), 3

Baron, E., Hauschildt, P. H., Branch, D., et al. 1995, ApJ, 441, 170

Baron, E., Nugent, P. E., Branch, D., & Hauschildt, P. H. 2004, ApJ, 616, L91

Bellazzini, M. 2008, Mem. Soc. Astron. Ital. 79, 440

Benedict, G. F., McArthur, B. E., Feast, M. W., Barnes, T. G. Harrison, T. E., et al. 2007, ApJ, 79, 453

Biscardi, I., Raimondo, G., Cantiello, M., & Brocato, E. 2008, ApJ, 678, 168

Blakeslee, J. P., Lucey, J. R., Tonry, J. L., Hudson, M. J., Narayanan, V. K., & Barris, B. J. 2002, MNRAS, 330, 443

Blakeslee, J. P., Jordan, A., Mei, S., Cote, P., Ferrarese, L., et al. 2009, ApJ, 694, 556

Blakeslee, J. P., Cantiello, M., Mei, S., Cote, P., DeGraaff, R. B., Ferrarese, L., et al. 2010, ApJ, 724, 657

Bono, G. 2003, in Stellar Candles for the Extragalactic Distance Scale, Lecture Notes in Physics, Vol. 635, ed. D. Alloin, & W. Gieren (Heidelberg: Springer), 85

Braatz, J. A., Reid, M. J., Humphreys, E. M., Henkel, C., Condon, J. J., & Lo, K. Y. 2010, ApJ, 718, 657

Branch, D. 1998. ARAA, 36, 17

Brunthaler, A., Reid, M. J., Falcke, H., Greenhill, M. J., & Henkel, C. 2005, Science, 307, 1440

Buchler, J. R. 2009, in Stellar Pulsation, Challenges for Theory and Observation, ed. J. Guzik, & P. Bradley. AIP Conference Proceedings, Vol. 1170, 51. astro-ph/0907.1766

Capaccioli, M., della Valle, M., Rossino, L., & D'Onofrio, M. 1989, AJ, 97, 1622

Capaccioli, M., della Valle, M., D'Onofrio, M., & Rosino, L. 1990, ApJ, 360, 63

Caputo, F. 2008, Mem. Soc. Astron. Ital. 79, 453

Catelan, M., Pritzl, B. J., & Smith, H. A. 2004, ApJS, 154, 633

Ciardullo, R. 2003, in Stellar Candles for the Extragalactic Distance Scale, Lecture Notes in Physics, Vol. 635, ed. D. Alloin, & W. Gieren (Heidelberg: Springer), 243

Ciardullo, R. 2005, in Planetary Nebulae as Astronomical Tools, AIP Conference Proceedings, Vol. 804, ed. R. Szczerba, G. Stasinska, & S. K. Gorny (Melville: American Institute of Physics), 277

Ciardullo, R., Jacoby, G. H., Ford, H. C., & Neill, J. D. 1989, ApJ, 339, 53

Ciardullo, R., Feldmeier, J. J., Jacoby, G. H., Kuzio de Naray, R., Laychak, M. B., et al. 2002, ApJ, 577, 31

Contreras, C., Hamuy, M., Phillips, M. M., Folatelli, G., Suntzeff, N. B., et al. 2010, AJ, 139, 519

Cox, J. P. 1980, Theory of Stellar Pulsation Princeton: Princeton University Press

Da Costa, G. S., & Armandroff, T. E. 1990, AJ, 100, 162

Darling, J. 2011, ApJL, 732, 2

de Grijs, R. 2011, An Introduction to Distance Measurement in Astronomy (Chichester: Wiley) de Jong, J. T. A., Rix, H., Martin, N. F., Zucker, D. B., Dolphin, A. E., Bell, E. F., Belokurov, V., & Evans, N. W. 2008, AJ, 135, 1361

Della Valle, M., & Livio, M. 1995, ApJ, 452, 704

Dessart, L., & Hiller, D. J. 2005, A&A, 439, 671

di Benedetto, G. P. 2008, MNRAS, 390, 1762

Dolphin, A. E. 2002, MNRAS, 332, 91

Eastman, R. G., & Kirshner, R. P. 1989, ApJ, 347, 771

Eastman, R. G., Schmidt, B. P., & Kirshner, R. 1996, ApJ, 466, 911

Feast, M. W., & Catchpole, R. M. 1997, MNRAS, 286, L1

Fernie, J. D. 1969, PASP, 81, 707

Ferrarese, L., et al. 2000, ApJ, 529, 745

Ferrarese, L., Cote, P., & Jordan, A. 2003, ApJ, 612, 1261

Folatelli, G., Phillips, M. M., Burns, C. R., Contreras, C., Hamuy, M., et al. 2010, AJ, 139, 120

Forbes, D. A., Brodie, J. P., & Grillmair, C. J. 1997, AJ, 113, 1652

Fouque, P., Arriagada, P., Storm, J., Barnes, T. G., Nardetto, N., et al. 2007, A&A, 476, 73

Freedman, W. L., & Madore, B. F. 2010, ARAA, 48, 673

Freedman, W. L., Madore, B. F., Monson, A, Persson, S. E., Scowcroft, V., Seibert, M., & Rigby, J. 2012, ApJ, 758, 24

Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., et al. 2001, ApJ, 553, 47

Freedman, W. L., Madore, B. F., Rigby, J., Persson, S. E., & Sturch, L. 2008, ApJ, 679, 71

Gascoigne, S. C. B., & Shobbrook, R. R. 1978, PASA, 3, 285

Gilmozzi, R., & Della Valle, M. 2003, in Stellar Candles for the Extragalactic Distance Scale, Lecture Notes in Physics, Vol. 635, ed. D. Alloin, & W. Gieren (Heidelberg: Springer), 229

Giovanelli, R., Haynes, M. P., Herter, T., Vogt, N. P., da Costa, L. N., et al. 1997, AJ, 113, 53

Girardi, L., & Salaris, M. 2001, MNRAS, 323, 109

Gnedin, O. Y. & Ostriker, J. P. 1998, in Galactic Halos, ASP Conference Series, Vol. 136, ed. D. Zaritsky, 56

Grocholski, A. J., & Sarajedini, A. 2002, AJ, 123, 1603

Hamuy, M., Phillips, M. M., Suntzeff, N. B., Schommer, R. A., Maza, J., et al., 1996, AJ,112, 2398

Hanes, D. A. 1977, MNRAS, 180, 309

Hanes, D. A. 1979, A new determination of the Hubble Constant, MNRAS, 188, 901

Heck, A., & Caputo, F. 1999, Post-Hipparcos Cosmic Candles, A&SSL (Dordrecht: Kluwer)

Henkel, C., Wouterloot, J. G. A., & Bally, J. 1986, A&A, 155, 193

Herrnstein, J. R., Moran, J. M., Greenhill, L. J., Diamond, P. J., Inoue, M., et al. 1999, Nature, 400, 539

Hicken, M., Wood-Vasey, W. M., Blondin, S., Challis, P., Jha, S., et al. 2009, ApJ, 700, 1097

Hillebrandt, W., & Niemeyer, J. C. 2000, ARAA, 38, 191

Hodge, P. 1982, ARAA, 19, 357

Hubble, E. E. 1925, ApJ, 62, 409

Hubble, E. P. 1926, ApJ, 63, 236

Hubble, E. P. 1929, ApJ, 69, 103

Huchra, J. P. 1992, Science, 256, 321

Humphreys, E. M. L., Reid, M. J., Greenhill, L. J., Moran, J. M. & Argon, A. L. 2008, ApJ, 672, 800

Iben, I., & Renzini, A. 1983, ARAA, 21, 271

Jackson, N. 2007 Living Rev. Relativ. 10, 4

Jacoby, G. H., Branch, D., Ciardullo, R., Davies, R. L., Harris, W. E., et al. 1992, PASP, 104, 599

Jacoby, G. H., & Ciardullo, R. 1993, in Planetary Nebulae, Proceedings IAU Symposium, Vol. 155, ed. R. Weinberger, & A. Acker (Dordrecht/Boston: Kluwer), 503

Jameson, R. F. 1986, Vistas Astron., 29, 17

Jha, S., Kirshner, R. P., Challis, P., Garnavich, P. M., Matheson, T., et al. 2006, AJ, 131, 527

Jones, W. C., Ade, P. A. R., Bock, J. J., Bond, J. R., Borrill, L., et al. 2006, ApJ, 647, 823

Kirshner, R. P., & Kwan, J. 1974, ApJ, 193, 27

Koerwer, J. F. 2009, AJ, 138, 1

Komatsu, E., Dunkley, J., Nolta, M. R., Bennett, C. L., Gold, B., et al. 2009, ApJS, 180, 330

Kowal, C. T. 1968, AJ, 73, 1021

Leavitt, H. S. 1908, Ann. Harv. Coll. Obs., 60, 87

Leavitt, H. S., & Pickering, E. C. 1912, Harv. Coll. Obs. Circ., 173, 1

Lee, M. G., Freedman, W. L., & Madore, B. F. 1993, ApJ, 417, 553

Livio, M., Donahue, M., & Panagia, N. 1999, The Extragalactic Distance Scale, STScI Symposium, No. 10 (Cambridge: Cambridge University Press)

Lo, K. Y. 2005, ARAA, 43, 625

Loeb, A., Reid, M. J., & Falke, H. 2005, ApJ, 633, 894

Longmore, A. J., Fernley, J. A., & Jameson, R. F. 1986, MNRAS, 220, 279

Macri, L. M., Stanek, K. Z., Bersier, D., Greenhill, L. J., & Reid, M. J. 2006, ApJ, 652, 1133

Madore, B. F., & Freedman, W. L. 1991, PASP, 103, 933

Madore, B. F., & Freedman, W. L. 2012, ApJ, 744, 132

Madore, B. F., Mager, V., & Freedman, W. F. 2009, ApJ, 690, 389

Mager, V., Madore, B. F., & Freedman, W. F. 2008, ApJ, 689, 721

Maoz, E., Newman, J. A., Ferrarese, L., Stetson, P. B., Zepf, S. E., et al. 1999, Nature, 401, 351

McGaugh, S. S., Schombert, J. M., Bothun, G. D., & de Blok, W. J. G. 2000, ApJL, 533, L99

Monson, A., Freedman, W. L., Madore, B. F., et al. 2012, ApJ

Mould, J., & Sakai, S. 2008, ApJL, 686, L75

Mould, J., & Sakai, S. 2009a, ApJ, 694, 1331

Mould, J., & Sakai, S. 2009b, ApJ, 697, 996

Mould, J. R., Huchra, J. P., Freedman, W. L., Kennicutt, R. C., Ferrarese, L., et al. 2000, ApJ, 529, 786

Newman, J. A., Ferrarese, L., Stetson, P. B., Maoz, E., Zepf, S. E., et al. 2001, ApJ, 553, 562

Nolta, M., Dunkley, J., Hill, R. S., Hinshaw, G., Komatsu, E., et al. 2009, ApJS, 180, 296

Paczynski, B., & Stanek, K. Z. 1998, ApJL, 494, 219

Perryman, M. A. C. 2009, Astronomical Applications of Astromentry: Ten Years of Exploitation of the Hipparcos Satellite Data (Cambridge: Cambridge University Press)

Phillips, M. M. 1993, ApJL, 413, L105

Pierce, M. J., & Tully, R. B. 1988, ApJ, 330, 579

Pietrzynski, G., Gieren, W., & Udalski, A. 2003, AJ, 125, 2494

Pietrzynski, G., Gorski, M., Gieren, W., Laney, D., Udalski, A., & Ciechanowski, A. 2010, AJ, 140, 1038

Pskovskii, Y. P. 1984, Sov. AJ, 28, 658

Readhead, A. C. S., Mason, B. S., Contaldi, C. R., Pearson, T. J., Bond, J. R., et al. 2004, ApJ, 609, 498

Reichardt, C., Ade, P. A. R., Bock, J. J., Bond, J. R., Brevik, J. A., et al. 2009, ApJ, 694, 1200

Reid, M. J., Braatz, J. A., Condon, J. J., Greenhill, L. J., Henkel, C., et al. 2009, ApJ, 695, 287

Richler, T. 2003, in Stellar Candles for the Extragalactic Distance Scale, Lecture Notes in Physics, Vol. 635, ed. D. Alloin, & W. Gieren (Heidelberg: Springer), 281

Riess, A. G., Macri, L., Casertano, S., Sosey, M., Lampeitl, H., et al. 2009a, ApJ, 699, 539

Riess, A. G., Macri, L., Li, W., Lampeitl, H., Casertano, S., et al. 2009b, ApJS, 183, 109

Riess, A. G., Macri, L., Casertano, S., Lampeitl, H., Ferguson, H. C., Filippenko, A. V., Jha, S. W., Li, W., & Chornock, R. 2011, ApJ, 730, 119

Rizzi, L., Tully, R. B., Makarov, D., Makarova, L., Dolphin, A. E., et al. 2007, ApJ, 661, 81

Rowan-Robinson, M. 1985, The Cosmological Distance Ladder, Distance and Time in the Universe (New York: W. H. Freeman)

Saha, A., Sandage, A. R., Tammann, G. A., Labhardt, L., Macchetto, F. D., et al. 1999, ApJ, 522, 802

Sakai, S., Mould, J. R., Hughes, S. M. G., Huchra, J. P., Macri, L. M., et al. 2000, ApJ, 529, 698

Sandage, A. R. 1968, ApJL, 152, 149

Sandage, A. R., Saha, A., Tammann, G. A., Labhardt, L., Panagia, N.,& Macchetto, F. D. 1996, ApJL, 460, 15

Sandage, A. R., & Tammann, G. A. 1982, ApJ, 265, 339

Sandage, A. R., & Tammann, G. A. 1990, ApJ, 365, 1

Sandage, A. R., & Tammann, G. A. 2006, ARRA, 44, 93

Sarajedini, A., 1999, AJ, 118, 2321

Sarajedini, A., Grocholski, A. J., Levine, J., & Lada, E. 2002, AJ, 124, 2625

Scowcroft, V., Freedman, W. L., Madore, B. F., Monson, A. J., Persson, S. E., Seibert, M., Rigby, J. R., & Sturch, L. 2011, ApJ, 743, 76

Scowcroft, V., Freedman, W. L., Madore, B. F., Monson, A., Persson, E., Seibert, M., Rigby, J., Stetson, P., & Sturch, L. 2012, ApJ, 747, 84

Schaeffer, B. 2008, AJ, 135, 112

Seibert, M., Freedman, W. L., Madore, B. F., Monson, A., Persson, E., Scowcroft, V. M., Rigby, J., Stetson, P., & Sturch, L. 2012, ApJ

Seidel, E., Da Costa, G. S., & Demarque, P. 1987 ApJ, 313, 192

Smith, R. 1982, The Expanding Universe, Astronomy's Great Debate 1900–1931 (Cambridge: Cambridge University Press

Smith, H. A. 1995, RR Lyrae Stars, Cambridge Astrophysics Series, Vol. 27 (Cambridge/New York: Cambridge University Press)

Soszynski, I., Poleski, R., Udalski, A., Szymanski, M. K., Kubiak, M., et al. 2008, Acta. Astron. 58, 163

Stanek, K. Z., Zaritsky, D. & Harris, J. 1998, ApJ, 500, L141

Steinmetz, M., & Navarro, J. 1999, ApJ, 513, 555

Tammann, G. A., & Sandage, A. 1999, in Harmonizing Cosmic Distance Scales in a Post-Hipparcos Era, ASP Conference Series, Vol. 167, ed. D. Egret, & A. Heck (San Francisco : Astronomical Society of the Pacific), 204

Tammann, G. A., Sandage, A. R., & Reindl, B. 2008, A&AR, 15, 289

Tolstoy, E., & Saha, A. 1996, ApJ, 462, 672

Tonry, J., & Schneider, D. P. 1988, AJ, 96, 807

Tonry, T. J., Dressler, A., Blakeslee, J. P., Ajhar, E. A., Fletcher, A. B., Luppino, G. A., Metzger, M. R., & Moore, C. B. 2001, ApJ, 546, 681

Tonry, T. J., Blakeslee, J. P., Ajhar, E. A., & Dressler, A. 2002, ApJ, 530, 625

Tully, R. B., & Fisher, J. R. 1977, A&A, 54, 661

Tully, R. B., & Pierce, M. J. 2000, ApJ, 553, 744

Udalski, A., 2000, Acta Astron., 50, 279

Udalski, A., Szymanski, M., Kubiak, M., et al. 1998, Acta Astron., 48, 147

van den Bergh, S. 1992, PASP, 104, 861

van den Bergh, S., & Pritchet, C. (eds.) 1988, The Extragalactic Distance Scale, ASP Conference Series, Vol. 4. (San Francisco: Astronomical Society of the Pacific)

Webb, S. 1999, Measuring the Universe (Praxis: Springer)

Weinberg, M. D., & Nikolaev, S. 2001, ApJ, 548, 712

Whelan, J., & Iben, I. J. 1973, ApJ, 186, 100

Wilcots, E. M., & Miller, B. W. 1998, AJ, 116, 2363

Wood-Vasey, W. M., Friedman, A. S., Bloom, J. S., Hicken, M., Modjaz, M., et al. 2008, ApJ, 689, 377

Zwicky, F. 1336, PASP, 48, 191

10 Galaxies in the Cosmological Context

Gabriella De Lucia
INAF – Astronomical Observatory of Trieste, Trieste, Italy

1 Introduction 453

2 The Framework: The Dissipationless Universe 455
2.1 The Cosmological Model 455
2.2 The Halo and Subhalo Mass Functions 456
2.3 Halo Structure 459
2.4 Halo Merger Trees 460

3 The Physics of Galaxy Formation 464
3.1 Gas Accretion 464
3.2 Gas Cooling 465
3.3 Star Formation 467
3.3.1 The First Generation of Stars 468
3.3.2 A Star Formation Law 468
3.3.3 The Initial Mass Function 470
3.4 Feedback 471
3.4.1 Photoionization Heating 472
3.4.2 Supernovae Feedback 473
3.4.3 AGN Feedback 474
3.5 Chemical Enrichment 477
3.6 Galaxy-Galaxy Interactions 477
3.7 The Environment 481
3.7.1 Galaxy Harassment 482
3.7.2 Cannibalism 482
3.7.3 Ram-Pressure Stripping 483
3.7.4 Strangulation 486
3.8 Stellar Populations 486

4 Putting It Together: Models of Galaxy Formation in a Cosmological Context . . . 487
4.1 The Halo Occupation Distribution Method 488
4.2 Hydrodynamical Simulations 490
4.3 Semianalytic Models of Galaxy Formation 492
4.4 Successes and Problems of Semianalytic Models of Galaxy Formation 493

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_10, © Springer Science+Business Media Dordrecht 2013

5 Concluding Remarks *498*

Acknowledgments *498*

References *499*

Abstract: In the last decades, a number of observational experiments have converged to establish the cold dark matter model as the "de facto" standard model for structure formation. While the cosmological paradigm appears to be firmly established, a theory of galaxy formation remains elusive, and our understanding of the physical processes that determine the observed variety of galaxy properties and their evolution as a function of cosmic time and environment is far from complete. Although much progress has been made, both on the theoretical and observational side, understanding how galaxies form and evolve remains one of the most outstanding questions of modern astrophysics. This chapter provides an introduction to ideas and concepts that underpin modern models of galaxy formation and evolution, in the currently favoured cosmological context.

1 Introduction

It was not until the seventeenth century that Galileo discovered that the swathe of light visible on a dark night from horizon to horizon was not made up of some sort of "celestial fluid" but was instead composed of myriads of unresolved stars. More and more "patches of light" started to be observed – *nebulae* or *Island Universes*, using the definition given by Immanuel Kant. A comet hunter – Charles Messier – and a musician who become a skilled maker of the most powerful telescopes of his time – Wilhelm Herschel – independently produced the first catalogs of nebulae. The designations introduced by Messier, for basically all the nebulae that can be seen with small telescopes, are still in use today (e.g., the nearest spiral galaxy to the Milky Way – Andromeda – is also known as M31). Despite a *Great Debate*[1] held in 1920 to establish the nature of these objects, the controversy remained unresolved until 1925 when Edwin Hubble, using distances estimated from Cepheid variables in M31, provided the definitive demonstration of their extragalactic nature. Since then, astronomers have made huge progress in the observation of extragalactic systems and have collected a vast amount of detailed information, in different portions of the electromagnetic spectrum, for millions of galaxies. Despite almost one century having passed since the birth of extra-galactic astronomy, and despite much progress from both the observational and theoretical side having been made, many questions about the formation and the evolution of galaxies remain unanswered.

How do the nebulae form? And how do they evolve as a function of cosmic time and environment? The first detailed models for the formation of galaxies were proposed only about 40 years after the confirmation of their extragalactic nature. In their classical paper, Eggen et al. (1962) analyzed the properties and motion of 221 dwarf stars and showed that those of lower metallicity tended to move on more highly eccentric orbits. The observed trends were interpreted as a signature that the stars that are observed as a spheroidal halo in our galaxy formed during a rapid radial collapse that later continued to form the stellar disk. This scenario was later worked out in more detail in early numerical simulations carried out by Larson (1975, 1976). His work showed that, with appropriate choices of the parameters, these dissipative

[1] The National Academy of Sciences in Washington invited two astronomers, Harlow Shapley and Heber Curtis, to "debate" about the scale of the universe and the nature of the *nebulae*. The debate had no winner or looser. Although Curtis turned out to be correct as he believed that the nebulae were galaxies external to our own, Shapley was correct in arguing that our galaxy was larger than previously thought and for showing that our Sun was not at the center of its galaxy.

collapse models can reproduce the observed basic properties of both elliptical and spiral galaxies, provided that the star formation is much slower in proto-spirals than in proto-ellipticals. The numerical work by Larson, however, also pointed out that if some means of redistributing angular momentum is not included (e.g., viscosity), these models are unable to obtain the high-surface brightnesses that are observed in real galaxies. We now know that one of the main problems with these early studies was that they neglected the presence of dark matter.

The first observational evidence of a *missing mass* problem dates back to the 1930s, when Zwicky (1937) estimated that the speeds of galaxies in the Coma cluster are too large to keep the system gravitationally bound, unless the dynamical mass is at least 100 times larger than the mass contained in galaxies. The reality of the problem, however, gained a hold upon the astronomical community only in the mid-1970s, when different studies showed that the rotation curves of spiral galaxies are either flat or rising at the optical edge of the galaxies, contrary to the Keplerian fall off that is expected if the visible stars and gas were the only mass in the system (Rubin and Ford 1970; Einasto et al. 1974; Ostriker et al. 1974). These observations led to the conclusion that dark matter must play an important role in galaxy formation, and motivated the two-stage theory proposed by White and Rees (1978). In this scenario, dark matter haloes form first, and the physical properties of galaxies are then determined by cooling and condensation of gas within the potential well of the haloes. This model contains many of the ideas that are at the basis of the tools that are nowadays used to study the formation and evolution of galaxies, and that will be discussed in more detail in this chapter.

In the 1980s, much work focused on the nature of the unseen dark matter component. Initially, many studies focused on neutrinos as the most likely candidates for the dark matter. It was soon realized, however, that in a neutrino-dominated universe, structure would form by fragmentation (top-down), with the largest superclusters forming first in a sort of flat "pancake"-like sheets (Zeldovich et al. 1982). These must then fragment to form smaller structures like galaxy groups and galaxies – a picture that conflicts with observation, as shown by detailed simulations of structure formation (White et al. 1983). During the same years, a number of different dark matter candidates were provided by particle physics models based on supersymmetry. These weakly interacting massive particles (WIMPS) are today considered the most likely candidates for dark matter. Because their masses are much larger (and therefore their velocities[2] are much smaller) than those of neutrinos, these particles are said to be "cold." Cold dark matter (CDM) decouples from the radiation field long before recombination so that its density fluctuations can grow significantly before the baryons decouple from the radiation. When this happens, baryons are free to fall in the dark matter potential wells that have formed and that allow structure formation to occur at a rate sufficient to be consistent with the large-scale structure observed at present (Davis et al. 1985). The CDM theory has now become the preferred scenario for galaxy formation and is the framework that will be adopted in this chapter. In a CDM universe, structure grows hierarchically (bottom-up), with small objects collapsing first and later merging in a continuous hierarchy to form more and more massive systems.

The aim of this chapter is to provide an introduction to the ideas and concepts that underpin modern models of galaxy formation and evolution, in a universe in which cosmic structures originate from small initial perturbations and build up hierarchically through gravitational instabilities. The layout of this chapter is as follows: ❯ Section 2 provides a brief description of the cosmological model that is currently accepted as the standard model for structure formation, while ❯ Sect. 3 deals with the physical processes that govern the formation and the

[2] Their velocities are nonrelativistic at the epoch of radiation-matter equality.

evolution of galaxies. ➲ Section 4 provides a brief review of the numerical techniques that are currently used to study galaxy formation in a cosmological context and highlights their most recent successes and open problems. Finally, ➲ Sect. 5 gives some concluding remarks. Due to space limits, this chapter does not contain a detailed overview of the observational properties of local and/or distant galaxies. The interested reader is referred to other chapters of this volume, as well as to the textbooks by, e.g., Binney and Merrifield (1998) and Mo et al. (2010), where also a more detailed exposition of some of the topics discussed in the following can be found.

2 The Framework: The Dissipationless Universe

This section provides a brief review of the cosmological framework in which galaxy formation and evolution take place, focusing on those ingredients that can be considered as the initial and boundary conditions for any galaxy formation model. For a more rigorous and detailed exposition of the subject, the reader is referred to classical textbooks by, e.g., Padmanabhan (1993) and Peacock (1999).

2.1 The Cosmological Model

During the last decade, a variety of observational tests have ushered in a new era of "precision cosmology" and have converged to establish the CDM model (Peebles 1982; Blumenthal et al. 1984) as the *de facto* standard cosmological model for structure formation. In the currently favored cosmogony (the ΛCDM universe), about 75% of the energy density is due to a yet unknown form of *dark energy* that tends to increase the rate of expansion of the universe, about 21% to a nonbaryonic *cold dark matter* that has yet to be detected in the laboratory, and only about 4% is made of baryonic matter out of which stars and galaxies are made. In the past years, it has been shown that this cosmological model is able to match simultaneously a variety of observational measurements, among which are the power spectrum of low-redshift galaxies, the structure that is seen in the Lyman α forest at $z \sim 3$, the present acceleration of the cosmic expansion as inferred from supernovae observations, and the temperature fluctuations in the cosmic microwave background. By combining these experiments, the parameters of this cosmological model are currently known with uncertainties of only a few percent (e.g., Komatsu et al. 2011), thus effectively removing a large part of the parameter space in galaxy formation studies.

The initial fluctuations are assumed to follow a Gaussian random distribution and to have expanded to cosmological scales by inflation[3] – a brief period of time during which the scale factor of the early universe increased exponentially. The dark matter component that has no pressure undergoes gravitational collapse, which makes the perturbations grow. The early evolution of these perturbations can be accurately described using the linear approximation which breaks down, however, when the density contrast becomes nearly unity. In the nonlinear regime, the evolution can be studied analytically if some simplifying assumptions are made (e.g., the

[3] The inflationary hypothesis was introduced by Guth (1981). While inflation is understood principally by its detailed predictions of the initial conditions for the hot early universe, the detailed particle physics mechanism responsible for it is not known.

spherical top-hat model, see, e.g., ➋ Chap. 8 of Padmanabhan) or, more directly and accounting for the full geometrical complexity of the problem, using cosmological N-body simulations.

For the purposes of modeling galaxy formation, the following information should be available: (i) the distribution of the dark matter halo masses at any given redshift, (ii) the structural properties of the dark matter haloes, and (iii) a statistical representation of their assembly history (that is what in the jargon is called a "merger tree").

2.2 The Halo and Subhalo Mass Functions

The first calculation of the abundance of gravitationally bound structures was carried out by Press and Schechter (1974), long before the CDM model was introduced. By assuming a Gaussian density field smoothed using a spherical top-hat window, and by varying the radius of the smoothing window, one can consider structures of different mass $M = 4/3\pi\rho R^3$. The abundance of haloes above a given mass depends on the fraction of spheres for which the density contrast (this is usually expressed as $\delta = \rho(x)/\bar{\rho} - 1$) exceeds some critical value δ_c. A natural choice for the critical value of the density contrast is provided by the spherical top-hat model and corresponds to the linearly extrapolated density contrast at which haloes are expected to virialize ($\delta_c \sim 1.69$).

Assuming that the probability that $\delta > \delta_c$ is the same as the fraction of mass elements that are contained in haloes with mass larger than M, one obtains

$$\frac{dn}{dM}(M,t) = \left(\frac{2}{\pi}\right)^{1/2} \frac{\rho_0}{M^2} \frac{\delta_c}{\sigma(M)} \left|\frac{d\ln\sigma}{d\ln M}\right| \exp\left[-\frac{\delta_c^2}{2\sigma^2(M)}\right] \tag{10.1}$$

where ρ_0 is the mean density of the universe, $\sigma(M)$ is the fractional root variance in the density field smoothed using a top-hat filter that contains, on average, a mass M, and $\delta_c(t)$ is the critical overdensity for spherical top-hat collapse at time t. The Press and Schechter derivation neglects underdense regions that can be enclosed within larger overdense regions and that would have a finite probability of being included in a larger collapsed object. To correct this, Press and Schechter introduced a "fudge factor" equal to 2 in front of the derived expression (this is included in the equation above) but did not give a proper demonstration of the correction adopted. An alternative derivation of the halo mass function was given by Bond et al. (1991), using what is usually referred to as the "excursion set formalism." A detailed expositionl of this formalism can be found in White (1994, see also Sect. 7.2 of Mo et al. 2010).

The halo mass function predicted by this simple calculation agrees surprisingly well with the results obtained from N-body simulations. This is shown in the left panel of ➋ *Fig. 10-1*. The colored symbols are results from the Millennium Simulation, which follows the evolution of $N = 2{,}160^3$ particles of mass $8.6 \times 10^8\ h^{-1}\mathrm{M}_\odot$ within a comoving box of size 500 h^{-1}Mpc on a side. Dashed lines are the Press-Schechter predictions at $z = 0$ and $z = 10$ and show that this formula underpredicts the high-mass end of the mass function by up to an order of magnitude, with the disagreement becoming worse at earlier cosmic epochs. Solid lines are predictions from the fitting formula proposed by Jenkins et al. (2001) that appears to describe results from the N-body simulation remarkably well, over the redshift and mass range well sampled.

Until the late 1990s, dissipationless simulations suffered from the so-called "overmerging" problem, i.e., substructures disrupted very quickly within dense environments so that haloes were smooth and featureless. The problem was initially explained by the lack of dissipation in N-body simulations (Katz et al. 1992; Summers et al. 1995): it was thought that baryons would

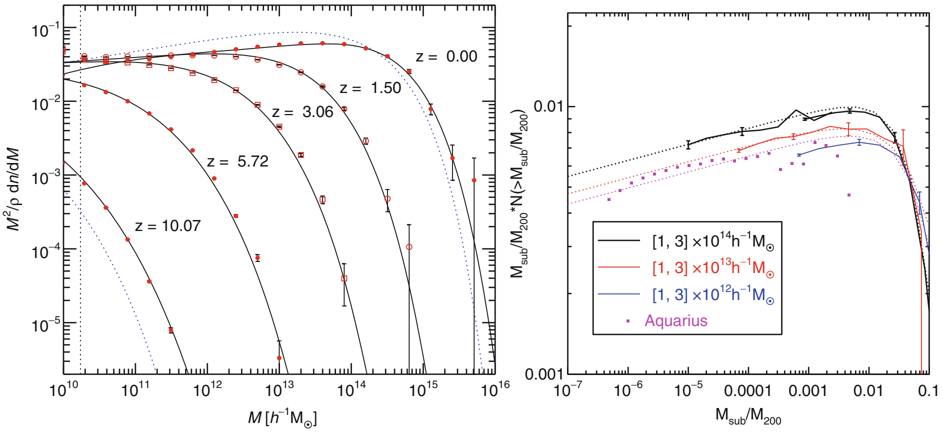


Fig. 10-1

***Left panel*: from Springel et al. (2005b). The differential halo mass function at different epochs. The mass function has been multiplied by M^2 to take out the dominant mass dependence. *Solid lines* are predictions from an analytic fitting function proposed in Jenkins et al. (2001) while *dashed lines* show the Press-Schechter mass function at $z = 0$ and $z \sim 10$. *Colored* symbols are obtained from the Millennium Simulation (Springel et al. 2005b). *Right panel*: from Gao et al. (2011). *Solid lines* show the averaged cumulative subhalo mass functions for three intervals of host halo mass as indicated in the legend. The error bars on selected points show the error on the mean for the three mass ranges indicated. The *filled squares* show the mean of the cumulative subhalo mass functions of the six Aquarius haloes with typical mass of ~10^{12} h^{-1}M$_\odot$**

sink into the center of dark matter haloes making them more resilient to disruption by the tidal field of the parent halo. Both analytic work and high-resolution simulations, however, demonstrated later that the cores of dark matter haloes that fall into a larger system can actually survive as self-gravitating objects orbiting in the smooth dark matter background of the halo, provided high enough force and mass resolution are used. A wealth of dark matter substructures are now routinely identified using different techniques (see below). If any, we are now facing the opposite problem of having “too much” substructure, at least on the galactic scale, where simulations predict more substructures than visible galaxies by almost two orders of magnitude (see Sect. 4 in Tasitsiomi 2003). As an example of the performance achieved by numerical N-body simulationsh in the last years, ➋ *Fig. 10-2* reproduces the density map of a 10-year-old high-resolution simulation of a galaxy cluster in the right panel, and what could be considered the state-of-the-art numerical simulation on the same scale only about three decades ago in the left panel.

The identification of substructures in dark matter haloes is a difficult technical problem, and many different algorithms have been developed to accomplish this task in the last years. Each of these has its own advantages and weaknesses. For example, in the hierarchical friends-of-friends algorithm (HFOF, Klypin et al. 1999), the linking length[4] is reduced in discrete

[4] The friend-of-friend (FOF, Davis et al. 1985) algorithm is a percolation algorithm that links together all the particles with a separation less than b times the mean interparticle separation. It has been shown that, with an appropriate choice of the linking length, it is possible to select groups close to the virial overdensity predicted by the spherical collapse model.

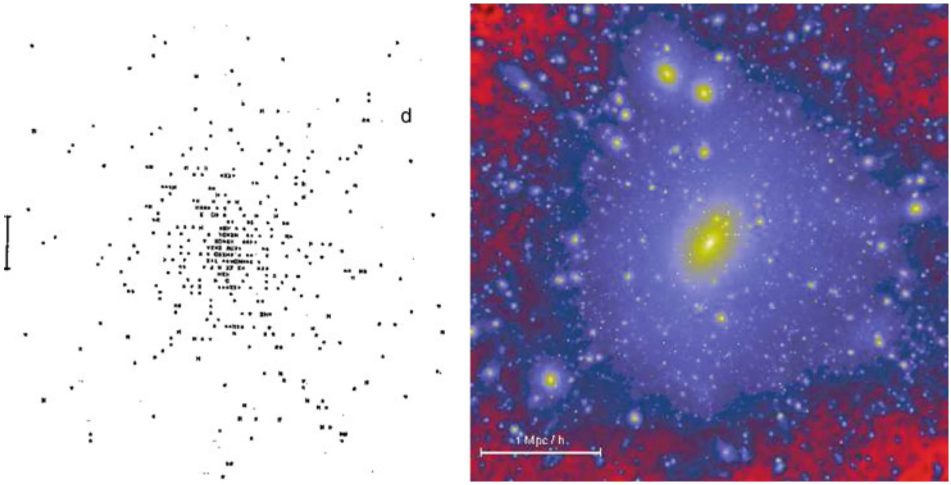


Fig. 10-2
***Left panel*: from White (1976). Projected distribution of a 700-body system with mass comparable to the virial mass of the Coma cluster. *Right panel*: from Springel et al. (2001). Density map of a high-resolution cluster resimulation. The cluster has a virial mass of 8.4×10^{14} $M_\odot$ and the high-resolution region of the simulation contains about 66 million particles**

steps, thus selecting groups of higher and higher overdensity. The choice of the levels of linking lengths is somewhat arbitrary, and the algorithm requires an iterative procedure. The bound density maximum algorithm (BDM, Klypin et al. 1999) iteratively determines a bound subset of particles in a sphere around a local density maximum. Since this method separates background particles from particles that are bound to the halo, the BDM algorithm estimates the physical properties of substructures more accurately. It implicitly assumes, however, that the halo is spherically symmetric, while the HFOF algorithm can deal with haloes of arbitrary shapes. Another approach is given by the SKID algorithm (see http ref: http://www-hpcc.astro.washington.edu/tools) in which the density around each particle is evaluated using a smoothing kernel. The particles are then moved along the density gradients toward a local density maximum. Particles that end up in the same local maximum are linked together using an FOF algorithm and then checked for self-boundness. Only self-bound groups with more than a user-specified minimum number of particles are kept as genuine substructures. An algorithm that has been frequently used in recent years is, finally, SUBFIND (Springel et al. 2001) which combines ideas used in other group-finding techniques with a topological approach for finding substructure candidates.

Typically, only about 10% of the total mass of a dark matter halo is found in substructures. The abundance of relatively massive substructures increases systematically (albeit weakly) with host halo mass, as shown in the right panel of *Fig. 10-1*. This trend reflects the fact that more massive haloes are both less centrally concentrated and younger (i.e., they assembled later) than their less massive counterparts. Therefore, they exert weaker tidal forces and have had less time to disrupt their substructures. As discussed above, different algorithms can be used to identify dark matter substructures, and different criteria for defining the boundaries and membership of these substructures are bound to lead to systematic differences. Several recent studies, however, find very similar slopes for the subhalo mass function, which suggests that the expected differences can be probably corrected by simple scale factors (a recent detailed comparison between different algorithms is given in Knebe et al. 2011).

2.3 Halo Structure

The internal structure of dark matter haloes has been studied extensively using N-body simulations. These show that the density profiles of dark matter haloes are shallower than r^{-2} at small radii and steeper at large radii. The density profile extracted from N-body simulations is well described by the following equation:

$$\rho(r) = \rho_{\rm crit} \frac{\delta_{\rm char}}{(r/r_s)(1+r/r_s)^2} \tag{10.2}$$

where r_s is a scale radius and $\delta_{\rm char}$ is a characteristic overdensity. The above profile has been shown to provide a good representation of the equilibrium density profiles of dark matter haloes of all masses in all CDM-like cosmogonies (Navarro et al. 1997, NFW). In (➋ 10.2), the local logarithmic slope gradually changes from a value of −3 in the outer parts to an asymptotic slope of −1 in the inner parts. The spatial scale r_s of this transition is treated as a fitting parameter and is often parameterized in terms of the concentration $c = r_h/r_s$ of the halo, which in fact is a reparametrization of $\delta_{\rm char}$ relative to the critical density:

$$\delta_{\rm char} = \frac{\Delta_h}{3} \frac{c^3}{\ln(1+c) - c/(1+c)}$$

where the limiting radius of a dark matter halo (r_h) is defined as the radius within which the mean matter density is

$$\rho_h = \Delta_h \bar{\rho} = \Delta_h \rho_{\rm crit} \Omega_m$$

and $\bar{\rho}$ is the mean matter density of the universe at the time considered and $\rho_{\rm crit}$ is the corresponding critical density for closure. Different definitions of the radius of a halo can be found in the literature. The most commonly adopted definition corresponds to R_{200}, that is, the radius that contains a mean overdensity equal to 200 times the critical density at the redshift considered. The corresponding enclosed mass is usually referred to as M_{200}, and in this case, $\Delta_h = 200/\Omega_m$.

For a given cosmology, the NFW profile is then completely characterized by the halo mass and by the concentration parameter $c = r_h/r_s$. At any given epoch, less massive haloes are more concentrated than their more massive counterparts (Neto et al. 2007 and references therein), a finding that can be interpreted as reflecting the density of the universe at the time of halo formation. More recent N-body studies (e.g., Navarro et al. 2004) show that the density profiles of highly resolved simulated haloes deviate from the NFW profile, particularly in the inner regions, and demonstrate that they are better described by an Einasto (1965) profile:

$$\rho(r) = \rho_{-2} \exp\left[\frac{-2}{\alpha}\left[\left(\frac{r}{r_{-2}}\right)^{\alpha} - 1\right]\right]$$

with r_{-2} equal to the radius at which the logarithmic slope of the density distribution is equal to −2 and $\rho_{-2} = \rho(r_{-2})$. The shape parameter of the Einasto profile (α) appears to vary systematically with halo mass (e.g., Hayashi and White 2008), a result that indicates a (small) deviation of the mean density profiles from a "universal" shape.

N-body simulations also show that dark matter haloes have strongly triaxial shapes, with a slight preference for nearly prolate systems (Jing and Suto 2002; Hayashi et al. 2007), and that they are supported by nearly isotropic velocity dispersions (Wojtak et al. 2005). Another important property of a dark matter halo is its angular momentum, traditionally parameterized as

$$\lambda = \frac{J\,E^{1/2}}{G\,M^{5/2}}$$

where J, E, and M are the total angular momentum, energy, and mass of the halo, respectively. Numerical simulations have shown that the distribution of spin parameters for dark matter haloes is well fit by a log-normal distribution:

$$p(\lambda)\lambda = \frac{1}{\sqrt{2\pi}\sigma_{\ln\lambda}}\exp\left[-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_{\ln\lambda}^2}\right]\frac{d\lambda}{\lambda}$$

with $\bar{\lambda} \sim 0.035$ and $\sigma_{\ln\lambda} \sim 0.5$. The median and width of this distribution appear to depend weakly on halo mass, redshift, and cosmology (Bett et al. 2007; Macciò et al. 2007).

2.4 Halo Merger Trees

A statistical description of the assembly history of haloes, i.e., a description of the merging events and of the masses of the haloes involved, can be obtained using a Monte Carlo approach by sampling the distribution of progenitor masses predicted from the extended Press-Schechter theory (Lacey and Cole 1993) or by using outputs from N-body simulations. In the jargon, this is called "merger tree." Its first schematic representation was presented by Lacey and Cole and is reproduced in ➔ *Fig. 10-3*. In the figure, cosmic time increases from top to bottom, and the widths of the branches reflect the masses of the individual merging haloes.

The excursion set approach of the extended Press-Schechter formalism provides a neat way to calculate the distribution of halo progenitor masses M_1 at redshift z_1, for a halo of mass M_2 at later redshift z_2. This can be written as

$$\frac{dN}{dM_1} = \left(\frac{2}{\pi}\right)^{1/2}\frac{d\ln\sigma}{d\ln M_1}M_2\frac{\sigma_1^2}{M_1^2}\frac{\delta_{c1}-\delta_{c2}}{(\sigma_1^2-\sigma_2^2)^{3/2}}\exp\left[-\frac{(\delta_{c1}-\delta_{c2})^2}{(\sigma_1^2-\sigma_2^2)}\right] \tag{10.3}$$

where $\sigma_1 = \sigma(M_1)$, $\sigma_2 = \sigma(M_2)$, $\delta_{c1} = \delta_c(z_1)$, $\delta_{c2} = \delta_c(z_2)$. Repeating the procedure at different redshifts, and imposing that the mass is conserved so that in each individual realization, the sum of the progenitor masses is equal to the mass of the parent halo, one can construct merger trees of haloes of different mass, with arbitrary high resolution. In practice, finding a suitable algorithm is not trivial, and different methods have been proposed (see Sect. 7.3 of Mo et al. 2010). In general, the Press-Schechter formalism and its Monte Carlo extension capture the qualitative behavior of all statistics that can be extracted from N-body simulations. However, recent studies have shown that some discrepancies are found between analytic merger trees and the corresponding statistics extracted from N-body simulations. The level of this disagreement, which becomes more important with increasing redshift, can be reduced by empirically tuning the progenitor distributions, but no theoretical justification exists for the form of the proposed corrections (Parkinson et al. 2008).

A fundamental assumption that underlies the Monte Carlo approach is that the formation history of a halo of a given mass does not depend on the "environment." This assumption

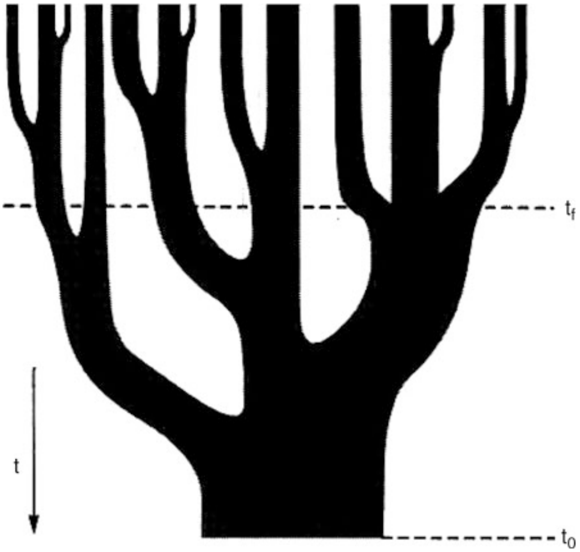


Fig. 10-3
From Lacey and Cole (1993). Illustration of a merger tree. Time increases from *top* to *bottom*, and the widths of the tree branches encode the masses of the merging haloes

was supported by early numerical work who found no dependence of halo clustering on concentration or formation time[5] (see, e.g., Lemson and Kauffmann 1999; Percival et al. 2003). A reanalysis of the same data, however, showed that close pairs of haloes form at slightly higher redshifts than more widely separated halo pairs, suggesting that haloes in dense regions form at slightly earlier times than haloes of the same mass in less dense regions (Sheth and Tormen 2004). These results were later confirmed by more recent numerical work that analyzed the properties of dark matter haloes in large volumes with high resolution, and found a clear dependency of the clustering amplitude on the halo formation time (Gao et al. 2005). This is illustrated in *Fig. 10-4* which shows the two-point correlation function[6] for haloes in four different mass ranges (each panel corresponds to a different mass bin, as indicated in the legend) and for the 20% oldest (red lines) and youngest (blue lines) haloes in each mass range. The figure shows that older haloes are more clustered than their younger counterparts with similar mass and that the dependence on the formation time is strongest for galactic mass haloes. It should be noted that these haloes were not well resolved by earlier numerical work that addressed the same issue.

Strictly speaking, this result invalidates the Monte Carlo approach in terms of using a one-parameter model (i.e., the halo mass) to construct the merger tree. In addition, as discussed above, this effect is strongest for haloes similar in mass to that of our Milky Way, which represent a large fraction of the galaxies in typical observational surveys. Since it is plausible that galaxy

[5] The formation time of a halo is typically defined as the time when the most massive progenitor of the halo first contains half the final mass.

[6] The two-point correlation function describes the probability, in excess of Poisson probability, to find two galaxies at a given relative distance.

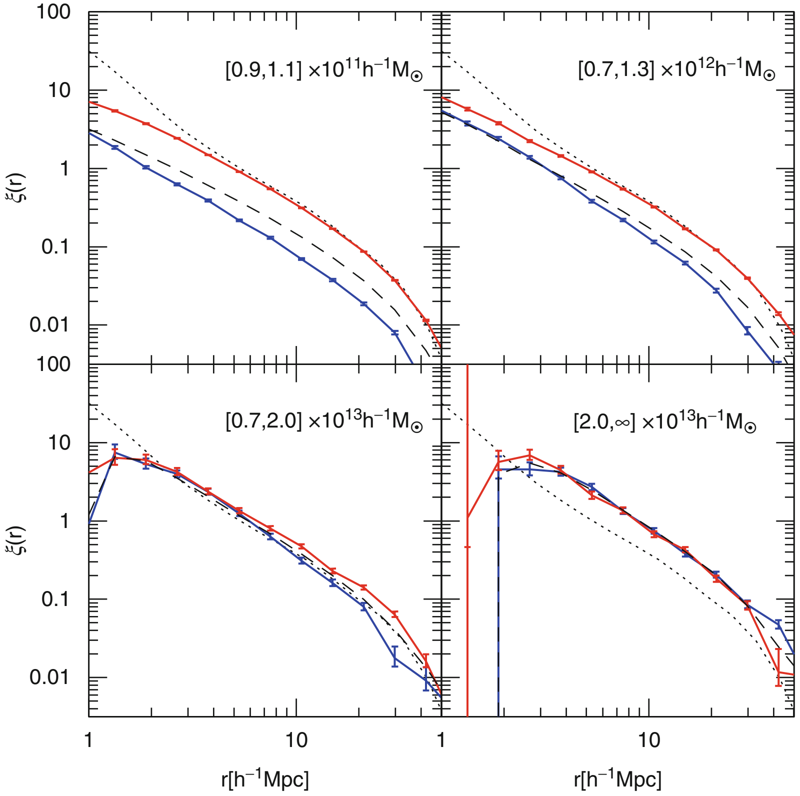


Fig. 10-4

From Gao et al. (2005). Two-point correlation functions for haloes in four mass ranges. Each panel gives results for haloes in the mass range indicated in the label. The *dotted black line*, repeated in all panels, is the correlation function of the underlying mass distribution. *Dashed black lines* give the correlation functions for the full sample of haloes in each mass range. The *red* and *blue* curves give correlation functions for the 20% oldest and 20% youngest of these haloes, respectively. Error bars are based on Poisson uncertainties in the pair counts

properties depend on the assembly history of their haloes, these results suggest that models that ignore the dependence on the large-scale structure will be in error, although the extent of the problem does likely depend on the specific galaxy formation model considered. Recent tests suggest that the effect discussed above influences the galaxy-galaxy and galaxy-mass correlation function by 5–10%, which is within their current statistical uncertainty (Croton et al. 2007). The trends discussed, however, are likely to play a more important role for studies of extreme objects that may be thought, for example, to form particularly early or late.

Alternatively, merger trees can be constructed using outputs from N-body simulations. This is not a trivial task: a discrete number of simulation outputs is available; one may want to include substructures which complicate significantly the merger tree structure; the mass of a halo can decrease with time; haloes may spatially overlap at a given time output and therefore be blended together by the specific group-finding algorithm employed, then separate at the next time output, and eventually come back together again later on; etc. (for a discussion of problems commonly encountered when building merger trees from N-body simulations, see Tweed et al. 2009). The main advantage of using merger trees extracted from simulations to graft on galaxy formation models is that they can give predictions for the positions of galaxies within haloes. This allows realistic "mock catalogues" to be constructed which contain not only the physical properties of all model galaxies (e.g., luminosities, masses, star formation rates) but also dynamically consistent redshift and spatial information, like in real galaxy redshift surveys. In addition, numerical merger trees are immune to the problem discussed above because they automatically take into account the dependence of halo clustering on age. On the other hand, N-body merger trees suffer of a finite mass resolution and of the "technical" problems mentioned above. Both approaches, extracting the trees directly from an N-body simulation and growing Monte Carlo trees, have therefore their advantages and weaknesses, and both are still widely used as input for some classes of galaxy formation models that will be discussed in more detail in Sect. 4.

Figure 10-5 shows the merger tree of a cluster-size halo, extracted from an N-body simulation. The branch highlighted in green is obtained by connecting the halo at each time step to the progenitor with the largest mass (the "main" progenitor). The rightmost branches are merger trees of secondary substructures (only those with more than 500 particles are shown) present in the FOF group at $z = 0$. Circles mark objects that belong to the same FOF group as the main progenitor, while triangles mark objects that have not yet joined the FOF group. Typically, when a halo is accreted onto a bigger system (i.e., becomes a "subhalo"), it loses mass efficiently due to tidal stripping (De Lucia et al. 2004a; Gao et al. 2004). A nice example of this process is

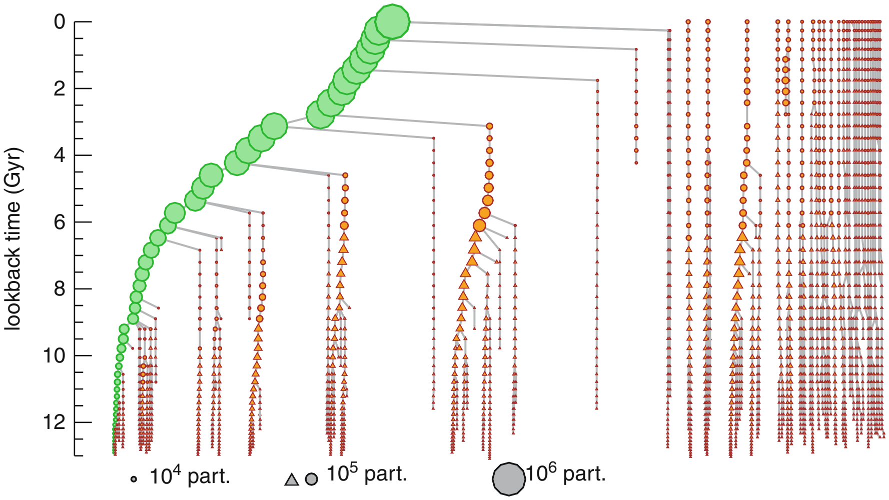


Fig. 10-5

From De Lucia and Blaizot (2007). Merger tree of a FOF group. Only the trees of subhalos with more than 500 particles at $z = 0$ are shown. Their progenitors are shown down to a 100-particle limit. The leftmost tree is that of the main subhalo of the FOF, while the trees on the right correspond to other substructures identified in the FOF group at $z = 0$

shown by the halo branch located roughly at the center of ➲ *Fig. 10-5*. The simulation work mentioned above also shows that a significant fraction of the substructures residing in cluster-size haloes at the present day were accreted at redshifts $z \lesssim 1$ and that the infall time and the retained mass of a subhalo are both strongly increasing functions of clustercentric radius. This implies that subhaloes in the inner regions of cluster haloes today were generally more massive in the past than similar mass but more recently accreted subhaloes in the outer regions. This is an important result to consider when linking the properties of luminous galaxies to those of dark matter (sub)haloes.

3 The Physics of Galaxy Formation

So far, this chapter has focused on the formation and evolution of structure under the influence of gravity alone. In order to make a close link between theoretical models of structure formation and observational data, it is necessary to consider the gas-dynamical and radiative processes that drive the evolution of the baryonic component of dark matter haloes. These processes are far more difficult to deal with than gravitational instability, as they cover several orders of magnitude in physical size and timescales, and are intertwined in an entangled network of actions, back reactions, and self-regulations. This section provides an overview of the main physical processes and ingredients that have to be considered when modeling the formation and evolution of galaxies in the cosmological set discussed in the previous section, highlighting the current status of observational and numerical studies.

3.1 Gas Accretion

During the linear regime, the density perturbation fields of the baryons and dark matter are expected to be equal on scales above the Jeans length. After halo formation, hydrodynamical forces come into play, and further collapse of the gaseous component associated with dark matter haloes is regulated by a combination of gravity, cooling, and hydrodynamical processes.

If the halo virial temperature exceeds the temperature of the accreting gas, then the gas will accrete supersonically, which will give rise to an accretion shock. Both analytic work and numerical simulations have early shown that when the cooling times are longer than the dynamical times, the shock occurs at a radius that is comparable (or slightly larger than) the virial radius (Bertschinger 1985; Evrard 1990). In reality, the accreting gas is not smooth but lumpy so that there is no well-defined accretion shock but rather a complex network of shocks. These heat the gas by thermalizing its kinetic energy up to the virial temperature of the halo. For an isothermal sphere, this can be written as

$$T_{\rm vir} = \frac{\mu m_{\rm p}}{2k_{\rm B}} V_{\rm c}^2 \simeq 3.6 \times 10^5 {\rm K} \left(\frac{V_{\rm c}}{100\,{\rm km\,s^{-1}}} \right)^2$$

where $m_{\rm p}$ is the proton mass and μ is the mean molecular weight of the gas. This gas will form a hydrostatically supported atmosphere which will obey the hydrostatic equilibrium equation:

$$\frac{{\rm d}P}{{\rm d}r} = \frac{{\rm d}(k_{\rm B} T \rho / \mu m_{\rm p})}{{\rm d}r} = -\rho(r) \frac{{\rm d}\Phi}{{\rm d}r} = -\rho(r) \frac{GM(r)}{r^2}$$

where P is the gas pressure, $\rho(r)$ the gas density, and $M(r)$ the total (i.e., dark matter plus baryonic) mass within the radius r. This gas will then cool radiatively and eventually lose energy and, consequently, pressure support. At this point, the gas will fall toward the center of the gravitational potential provided by the dark matter halo, conserving its angular momentum and settling in a denser gas disk.

In the regime where the cooling times are much shorter than the dynamical times, the shock forms at much smaller radii, close to the forming galaxy. The gas is still heated to very high temperatures (actually larger than in the slow cooling regime because the preshock velocity of the infalling gas will be larger than in this case of a virial shock) but will cool so rapidly that it cannot maintain the pressure needed to support a quasistatic hot atmosphere. The distinction between these "rapid" and "slow" cooling regimes was clearly understood when the first hierarchical galaxy formation models were presented (Rees and Ostriker 1977; Binney 1977; White and Frenk 1991). This picture has been validated by 1D hydrodynamical simulations (Birnboim and Dekel 2003, see also unpublished work by Forcada-Miro and White 1997) and by more recent 3D hydrodynamical simulations (Kereš et al. 2005; Ocvirk et al. 2008) that show that most of the accretion on haloes with mass $\lesssim 10^{12} \mathrm{M}_{\odot}$ tend to be directed along filaments, and is often referred to as "cold accretion." As Croton et al. (2006) have stressed and as noted above, the term cold accretion is a misnomer. In fact, what differentiates mainly the two modes of accretion is not the temperature to which infalling gas is shocked but rather the time spent by the gas at the postshock temperature before its energy is radiated away. It is worth noting that the transition mass between the rapid and slow cooling regimes found in the most recent simulations is very close to that identified in early analytical work (see discussion in Benson and Bower 2011). Finally, it should be noted that the rates computed in simulations often correspond to accretion rate onto the haloes and that these are different from the accretion rates onto the galaxies. The latter can be strongly affected by metal line cooling and by feedback from supernovae and/or active galactic nuclei (Benson and Bower 2011; van de Voort et al. 2011).

3.2 Gas Cooling

The primary cooling processes relevant for structure formation are two-body radiative processes. A gas with primordial composition (only hydrogen and helium) is almost entirely ionized at temperatures above 10^6 K, while a gas of nonzero metallicity is fully ionized at temperatures above a few 10^7 K. At these high temperatures, the cooling is dominated by the bremsstrahlung continuum due to the deceleration of electrons as they encounter atomic nuclei. At lower temperatures (i.e., $10^4\mathrm{K} < T < 10^6\mathrm{K}$), collisional ionization, recombination, and collisional excitation become important. At even lower temperatures ($T < 10^4$ K), most of the electrons have recombined so that atomic cooling is very inefficient. Cooling can still take place (albeit at very low rates) if the gas is enriched, but the dominant cooling in this regime is given by the excitation (through collisions) of rotational or vibrational energy levels of molecular hydrogen (or of other molecules if present) and subsequent decay. Since the dominant cooling processes are two-body processes, one can write the cooling rate per unit volume as

$$\mathcal{L} = n_{\mathrm{H}}^2 \Lambda(T, Z)$$

where n_{H} is the number density of hydrogen (both neutral and ionized) and $\Lambda(T, Z)$ is the cooling function that, as explained above, will depend (strongly) both on the temperature and on the chemical composition of the gas.

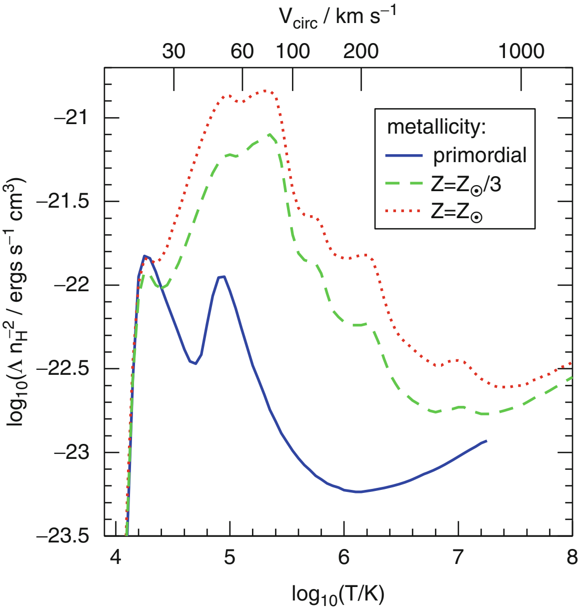


Fig. 10-6

From Baugh (2006), based on model results from Sutherland and Dopita (1993). The cooling rate is plotted as a function of the virial temperature of the hot halo gas. The equivalent circular velocity of the halo is indicated on the *top* axis. The different *curves* show how the cooling rate depends upon the metallicity of the gas, as indicated by the legend

Figure 10-6 shows how the cooling rate varies as a function of the temperature of the hot halo gas and how it depends upon the chemical composition of the gas. Cooling is dominated by bremsstrahlung at the high temperature end, where $\Lambda \propto T^{1/2}$. The peaks in the primordial cooling function at ~15,000 K and ~10^5 K are due to the collisionally excited electronic levels of hydrogen and singly ionized helium, respectively. For an enriched gas, cooling is significantly enhanced at temperatures $\gtrsim 10^5$ K due to the collisionally excited levels of ions of oxygen, carbon, nitrogen, etc. Above ~10^6 K, other metal lines contribute significantly, in particular neon, iron, and silicon. The cooling functions shown in *Fig. 10-6* are based on model results from Sutherland and Dopita (1993) and assume ionization equilibrium, i.e., that the densities of all ions are equal to their equilibrium values. This approximation is correct if the time scales of the radiative processes are much shorter than the hydrodynamical time scales of the gas, which might not be the case in shocks or in the case of a very dilute gas (where reaction rates are low). In these cases, a more appropriate treatment requires cooling rates to be recomputed using nonequilibrium densities.

At high redshifts, an additional cooling channel has to be taken into account: inverse Compton scattering of cosmic microwave background photons by electrons in the ionized plasma inside dark matter halos. This channel is effective if the temperature of the plasma exceeds that of the microwave background $T_\gamma \approx 2.73(1+z)$ K. It can be shown that

$$\frac{t_{\mathrm{Compton}}}{t_{\mathrm{age}}} \approx 350\Omega_m^{1/2} h(1+z)^{-5/2}$$

where $t_{\rm age}$ is the age of the universe at redshift z. For $\Omega_m = 0.3$ and $h = 0.7$, one obtains $t_{\rm Compton}/t_{\rm age} = 1$ at $z \sim 6$. So Compton cooling against the cosmic microwave background becomes important only at $z \gtrsim 6$. The cooling rate per unit mass associated with Compton cooling is proportional to the electron temperature and independent on the gas density (see Sect. 8.1.2 of Mo et al. 2010). So, assuming an isothermal distribution and a constant electron fraction, gas that is able to cool via this process will do so at all radii.

In a spherically symmetric gaseous system, a local cooling time can be defined dividing the thermal energy density of the gas by the cooling rate per unit volume:

$$t_{\rm cool}(r) = \frac{3}{2}\frac{kT\rho_{\rm g}(r)}{\bar{\mu}m_{\rm p}n_{\rm e}^2(r)\Lambda(T,Z)}$$

where $n_{\rm e}(r)$ is the electron density and $\rho_{\rm g}(r)$ is the gas density at a radius r. A simple estimate of the instantaneous cooling rate onto the central object can be obtained by following the method proposed by White and Frenk (1991): a cooling radius, $r_{\rm cool}$, can be defined as the radius at which the local cooling time is equal to a suitably defined age for the halo, e.g., the Hubble time. At early times and for low-mass haloes, the cooling radius can be larger than the virial radius. In this case, the hot gas is never expected to be in hydrostatic equilibrium, and the cooling rate is essentially limited by the accretion rate (the halo is in the rapid cooling regime discussed above). At late times and in massive systems, the cooling radius lies within the virial radius, and the gas can be assumed to cool quasistatically with a cooling rate that can be modeled by a simple inflow equation (this is the slow cooling regime):

$$\frac{{\rm d}M_{\rm cool}}{{\rm d}t} = 4\pi\rho_{\rm g}(r_{\rm cool})r_{\rm cool}^2\frac{{\rm d}r_{\rm cool}}{{\rm d}t}$$

The cooling model just described is extremely simplified and does not account, for example, for the fact that the gas distribution can readjust itself once gas starts to cool out (Viola et al. 2008). Nevertheless, it has been shown to provide results that are statistically in relatively good agreement with more detailed hydrodynamical simulations that adopt the same physics (e.g., Benson et al. 2001; Yoshida et al. 2002, but see also Saro et al. 2010). A number of assumptions need to be made, however, to implement the above simple prescriptions in analytic models of galaxy formation (e.g., about the gas profile and for the calculation of the cooling radius). Recent work has shown that the different assumptions adopted can give rise to significant differences, in particular at scales larger than those typical of our own galaxy (De Lucia et al. 2010).

3.3 Star Formation

It is generally accepted that the rate at which galaxies can form stars is determined by its ability to form dense molecular clouds. This is supported by direct observations of associations of young stars in the Milky Way and other nearby galaxies, as well as by observations of CO emission from starburst galaxies. From the theoretical point of view, that of star formation remains a poorly understood mechanism where processes like turbulence, magnetic fields, dust, molecular cooling, etc., all play an important role (a recent review can be found in McKee and Ostriker 2007). In terms of building a galaxy formation model, it is important to understand: (i) where and when the first generation of stars formed and their properties, (ii) the rate at which stars form in disks and starbursts, and (iii) the distribution of stellar masses produced in episodes of star formation.

3.3.1 The First Generation of Stars

If we believe that the structures in the universe grew hierarchically, the first objects that became nonlinear are expected to have masses much smaller than those of typical galaxies. The first generation of stars is expected to be extremely metal poor, because heavy elements can only be produced in the interior of stars. These stars are referred to as Population III stars. In a CDM model, the virial temperature of a halo is related to its mass through the following equation:

$$T_{\rm vir} \sim 442\Omega_m^{1/3}\left(\frac{M}{10^4 h^{-1}{\rm M}_\odot}\right)^{2/3}\left(\frac{1+z_{\rm vir}}{100}\right)K$$

where it has been assumed that the average density of dark matter haloes is 200 times the critical density, $H(z) = H_0\Omega_m^{1/2}(1+z)^3/2$ for $z >> 1$, and $z_{\rm vir}$ is the assembly redshift of the halo. At $z \gtrsim 200$, Compton scattering plays an important role, and the temperature of the intergalactic medium is

$$T_{\rm gas} = T_\gamma = 2.73\,(1+z)\,K$$

where T_γ is the temperature of the cosmic microwave background. At these redshifts, only haloes with $M \gtrsim 10^4{\rm M}_\odot$ can trap significant amount of baryonic gas. At lower redshift, the temperature of the intergalactic medium decreases faster than that of the microwave background, and lower mass haloes start trapping baryonic gas. The gas that is trapped will eventually be heated to the virial temperature of the parent haloes by shocks. If this gas can cool, it will condense and eventually fragment allowing the formation of stars in these early haloes.

As discussed in the previous section, at temperatures lower than ~10^4 K, the only significant cooling is due to molecular hydrogen, so the chemistry of this molecule governs the formation of the first objects. ➋ *Figure 10-7* shows the minimum mass of haloes within which H_2 cooling is sufficiently effective to lead to gas collapse (for a more rigorous exposition, see Ciardi and Ferrara 2005 and references therein). This minimum halo mass turns out to be between 10^4 ${\rm M}_\odot$ at $z \gtrsim 100$ and $10^8{\rm M}_\odot$ at lower redshift. In a CDM universe, haloes in this mass range start forming in large numbers only at $z \lesssim 30$. Significant uncertainties are involved in the discussion outlined above. For example, the presence of UV photons can dissociate hydrogen molecules and therefore suppress significantly the cooling efficiency. On the other hand, induced formation of hydrogen molecules behind shocks driven by the first stars can outweigh its photodissociation. In addition, once star formation begins, mechanical and radiative feedback from the first stars can remove a large fraction of the remaining halo gas.

Because of the very low metallicities, the cooling time of the gas may be significantly longer than the time scale of the gravitational collapse so that the cloud may not be able to fragment. First simulations of this initial collapse phase have confirmed this scenario and concluded that the first stars formed in isolation and were very massive (of the order of 60–100 ${\rm M}_\odot$ but with large uncertainties, see, e.g., Abel et al. 2002; Yoshida et al. 2006). More recent numerical studies have shown that metal-free gas clouds can fragment strongly, with the details of the process depending on the degree of turbulence in the halo. As a consequence, the mass spectrum of Pop III stars might be relatively flat ranging from ~0.1 to ~10 ${\rm M}_\odot$ (Clark et al. 2011; Greif et al. 2011).

3.3.2 A Star Formation Law

The problems related to the formation of the first stars propagate into galaxy formation theory if we wish to understand the rate at which stars form in these systems and any consequences that

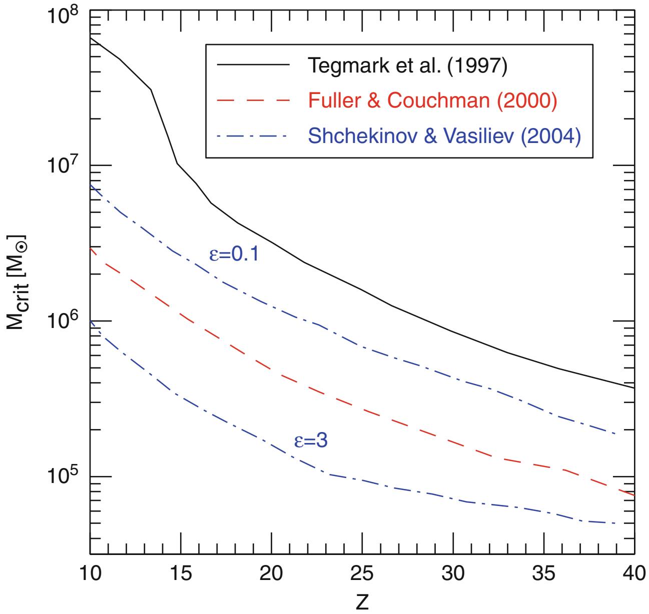


Fig. 10-7
From Ciardi and Ferrara (2005). Minimum mass able to cool and collapse as a function of redshift as calculated in the studies indicated in the legend. The *blue dot-dashed lines* are derived for two different values of the rate of ionizing photon production by ultra-high energy cosmic rays

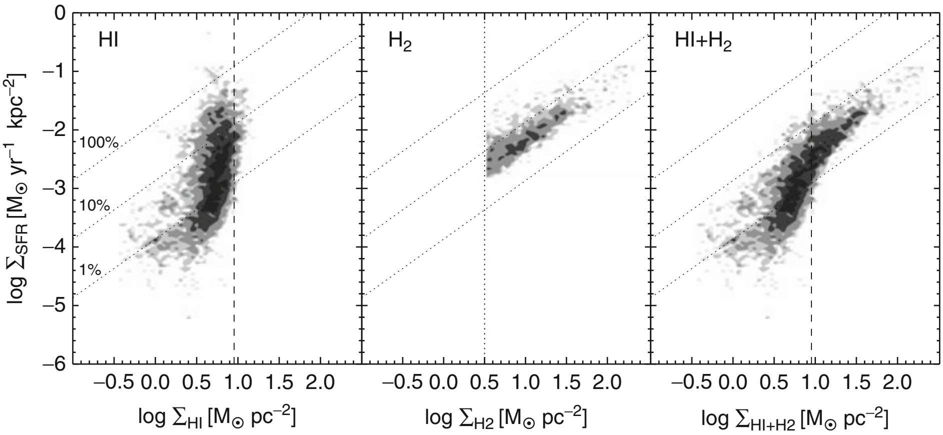
star formation might have for further evolution. Our limited understanding of the physical processes involved prevents us from constructing a "star formation law" from first principles that describes how the star formation rate Σ_{star} depends on the physical conditions of the interstellar medium. In order to make progress, one can appeal to empirical laws. A power-law relation of the form

$$\Sigma_{\text{star}} \propto \Sigma_{\text{gas}}^{N} \tag{10.4}$$

has been known since a long time (Schmidt 1959) and has been shown to provide a very good parametrization of the global star formation rate over a large range of surface densities, from the gas-poor spiral disks to the cores of the most luminous starburst galaxies (Kennicutt 1998). The best-fit observational data is

$$\Sigma_{\text{star}} = (2.5 \pm 0.7) \times 10^{-4} \left(\frac{\Sigma_{\text{gas}}}{M_{\odot \text{pc}^{-2}}} \right)^{1.4 \pm 0.15} M_\odot \, \text{year}^{-1} \text{kpc}^{-2}$$

where $\Sigma_{\text{gas}} = \Sigma_{HI} + \Sigma_{H2}$ is obtained averaging over the entire star-forming disk. Recently, it has become possible to study the star formation law by fitting Schmidt laws to individual galaxies for which Σ_{gas} and Σ_{star} are measured in azimuthally averaged rings or even on a pixel-by-pixel basis. ➋ *Figure 10-8* is based on a recent study by Bigiel et al. 2008 and shows the local star formation rate per unit area measured on a scale of ~750 pc as a function of the local atomic gas density (left panel), molecular gas density (middle panel), and total gas density (right panel). The atomic hydrogen distribution saturates at about $10 \, M_\odot pc^{-2}$, and the figure shows that it



■ Fig. 10-8
From Bigiel et al. (2010), based on data published in Bigiel et al. (2008). Local star formation rate per unit area (measured on a scale of ~750 pc from seven nearby spirals) as a function of the local atomic gas density (*left panel*), molecular gas density (*middle panel*), and total gas density (*right panel*). The *diagonal dotted lines* show lines of constant star formation efficiency, indicating the level of star formation needed to consume 1%, 10%, and 100% of the gas in 10^8 years. The *dashed vertical lines* in the *left* and *right* panels indicate the surface density at which HI saturates

poorly correlates with the star formation rate measured. Gas in excess of this value is predominantly molecular, and the middle panel of ❯ *Fig. 10-8* shows that there is a well-defined law for this gas component, which is well described by a power law with slope $N \sim 1$. This implies a constant molecular hydrogen depletion time of ~2 Gyr. As argued in Bigiel et al. (2008), the star formation law can be interpreted as a combination of two laws that regulate the conversion of atomic to molecular hydrogen and the formation of stars from molecular gas, respectively. Variants of ❯ 10.4 are commonly adopted in galaxy formation models where the formation of molecular gas is not usually followed explicitly. While this allows us to bypass the question of how stars form, it should be noted that such an empirical relation is then applied also beyond the regimes where it has been originally measured.

3.3.3 The Initial Mass Function

Galaxy properties depend not only on the rate and efficiency of star formation but also on the mass spectrum with which stars form, that is, the initial mass function (IMF). Observational results for our Milky Way suggest that the IMF has roughly the same form, independent of the location in the galaxy. The first determination of the IMF in the solar neighborhood was obtained by Salpeter (1955) who found that it is well described by a power law:

$$\phi(m)dm \propto m^{-b}dm$$

with $b = 2.35$, for stars in the mass range $0.4\,M_\odot \lesssim m \lesssim 10M_\odot$. $\phi(m)dm$ provides the relative number of stars born with masses in the range $m \pm dm/2$. Different measurements have been made more recently, and they suggest that the IMF deviates from a pure power law, becoming flatter at the lowest mass end and steeper at the highest mass end. All subsequent determinations do not deviate significantly from the Salpeter IMF for masses $\gtrsim 1\,M_\odot$, while for lower masses, there are significant differences among different determinations. One of the most recent measurements that is widely used in current years has been made by Chabrier (2003):

$$\xi(m) \propto \begin{cases} m^{-1.35} \text{ for } m > 1 M_\odot \\ \exp\left(-[\log(m/0.2M_\odot)]^2/0.6\right) \text{ for } m < 1M_\odot \end{cases}$$

where $\xi(m)$ is the logarithmic IMF and is defined as $\xi(m) d\log m = \phi(m) dm$.

The question of the "universality" of the IMF is a heavily debated one, particularly in recent years. From the observational point of view, there are large uncertainties, and only a small number of local star-forming clouds can be studied in detail (for a critical review of observational measurements, see Bastian et al. 2010). From the theoretical point of view, it is worth reminding that the ability of a gas cloud to collapse and fragment depends on the local Jeans mass:

$$M_J \simeq 700\ M_\odot (T/200\,\mathrm{K})^{3/2} (n/10^4\ \mathrm{cm}^{-3})^{-1/2} (\mu/2)^{-2},$$

where T, n, and μ are the temperature, number density, and mean molecular weight at the halt of fragmentation. For gas of primordial composition, a minimum temperature of ~200 K is reached when molecular hydrogen cooling becomes inefficient. This gives a Jeans mass $M_J \simeq 10^3\, M_\odot$. If metals are present, cooling can proceed to lower temperatures allowing the collapsing gas cloud to undergo fragmentation and form smaller clumps. The IMF can therefore depend on the metallicity: the hydrodynamical simulations by Smith and Sigurdsson (2007) show that above a critical metallicity, of about $10^{-3}\, Z_\odot$, clouds can fragment to form low-mass stars, while for gas of lower metallicities, stars form following a more top-heavy IMF. The critical metallicity defined above is well below that of the observed galaxies and, therefore, this effect might not be significant for galaxy formation studies. Further work is, however, needed to clarify if the IMF depends on the metallicity also above this critical value.

3.4 Feedback

The importance and the need of physical mechanisms that are able to modulate the efficiency of galaxy formation as a function of halo mass was recognized early on: Larson (1975, 1976) noted that supernovae-driven winds could remove most of the gas and heavy elements from low-mass galaxies. White and Rees (1978) argued that feedback is required to explain the overall low efficiency of galaxy formation. If dark matter haloes represent the birthplaces of luminous galaxies, evidence for the need of feedback comes from the observation that the shape of the halo mass function is very different from the shape of the observed luminosity function of galaxies. Thus, a simple model that assumes a fixed mass-to-light ratio would overpredict by order of magnitudes both the number of faint galaxies and that of bright ones (see left panel of ➋ *Fig. 10-9*). By matching the observed galaxy groups to dark matter haloes that are predicted to have the same space density, it is possible to derive the mass-to-light ratio that guarantees a match between theoretical predictions and the observed luminosity function (see right panel of ➋ *Fig. 10-9*). The required mass-to-light ratio is lowest for haloes of mass ~$10^{12}\, M_\odot$ which are, in other words, those in which galaxy formation is most efficient. In addition, this simple exercise confirms that the overall efficiency of galaxy formation must be low since most baryons do not end up as stars.

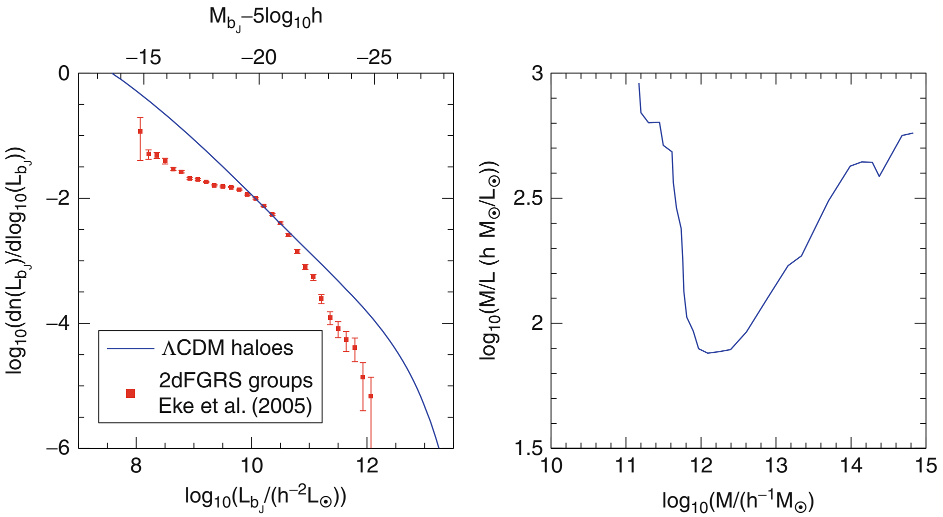


Fig. 10-9

From Baugh (2006). *Left panel*: the *solid line* has been obtained converting the dark matter halo mass function into a galaxy luminosity function by assuming a fixed mass-to-light ratio chosen to match the knee of the luminosity function. Data points with error bars show observational measurements. *Right panel*: the mass-to-light ratio required to match the observed luminosity function

We now know that feedback processes are those that have arguably the strongest influence on the observed galaxy properties but also those that are the most difficult to model. Broadly speaking, galaxy formation models consider three different forms of feedback: photoionization heating, supernovae feedback, and feedback from active galactic nuclei (AGN). The first two are believed to play an important role in shaping the faint end of the luminosity function, while the latter is believed to play a crucial role in regulating the condensation of gas in relatively massive haloes, thereby reducing the number of bright galaxies that would be predicted in the simple model outlined above. The following sections describe in more details these three modes of feedback and comment on recent theoretical results.

3.4.1 Photoionization Heating

It is believed that the hydrogen in the intergalactic medium must have been reionized somewhere between $z \sim 6$ and $z \sim 30$. Although it remains uncertain which energy sources were responsible for reionization, it was soon realized that the photoionizing background responsible for reionizing the intergalactic medium may also act to inhibit galaxy formation. In particular, this process acts in two different ways: (i) it heats the gas increasing its thermal pressure and therefore inhibiting its accretion onto dark matter haloes, and (ii) it also heats the gas that has already collapsed in haloes, therefore reducing the abundance of neutral atoms which can be collisionally excited, which in turn reduces the rate of radiative cooling of gas inside haloes (Doroshkevich et al. 1967; Efstathiou 1992). Both these mechanisms can effectively suppress galaxy formation in small haloes. ➲ *Figure 10-10* shows examples of net

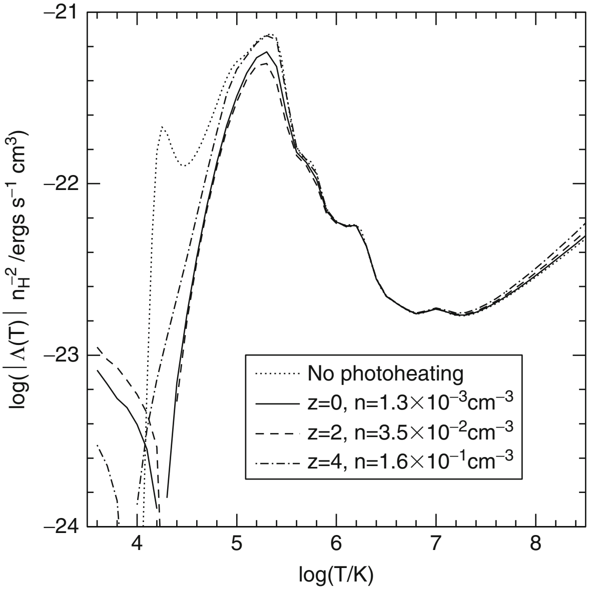


Fig. 10-10

From Benson et al. (2002). The net cooling/heating function for gas at different redshifts in the presence of the photoionizing background predicted in a fiducial model of galaxy formation. The absolute value of the cooling/heating rate is plotted per unit volume, for a gas with metallicity 0.3 $Z_\odot$

heating/cooling functions in the presence of a photoionizing background. These were calculated coupling the photoionization background computed self-consistently from a galaxy formation model with a photoionization code (for details, see Benson et al. 2002) and are computed for the typical densities of gas at each redshift indicated in the legend. The figure shows that photoionization can significantly suppress cooling in haloes with virial temperature in the range 10^4–10^5 K and therefore inhibit the formation of low-mass galaxies.

The value of the characteristic mass, M_c, below which galaxies are strongly affected by photoionization was calculated by Gnedin (2000) who argued that $M_c = M_F$, i.e., the filtering mass that corresponds to the scale over which baryonic perturbations are smoothed in linear perturbation theory. This relation has often been used in galaxy formation models to explain the low number of satellites observed in the Local Group. Recent numerical work has shown that the fitting function provided by Gnedin overestimates the characteristic mass by large factors (Okamoto et al. 2008).

3.4.2 Supernovae Feedback

The mechanical energy supplied by massive stars in the form of supernovae and stellar winds represents the engine that drives the galactic-scale outflows that are observed in the most actively star-forming galaxies both in the local universe and at high redshift. Observations suggest that outflows are ubiquitous in galaxies in which the global star formation rate per unit area exceeds roughly $10^{-1}M_\odot$ year^{-1} kpc^{-2} and that the material is multiphase containing cold, warm, and hot gas, plus dust and magnetized relativistic plasma. Different techniques and datasets can be used to estimate the mass and energy outflow rates. The available data suggest that the outflow rates are comparable to the star formation rates and that radiative losses in superwinds are not significant (Heckman 2002, and references therein). The estimated outflow speeds can be significant (in the range from hundreds to thousands km s^{-1}), and recent studies suggest that there is a weak trend with the galaxy star formation rate (Weiner et al. 2009, see also Martin 2005). This trend seems to support a picture in which winds are momentum driven through radiation pressure (Murray et al. 2005) rather than by the kinetic energy of supernova ejecta by entrainment in the hot wind (Strickland and Stevens 2000, and references therein). Unfortunately, the observational measurements available refer to material that is still relatively deep within the gravitational potential of the halo. So the estimated outflow rates are difficult to translate into rates at which mass, metals, and energy escape from galaxies and are eventually transported into the intergalactic medium. The fate of the winds (or superwinds depending on their velocity) will depend critically on a number of unknowns and on the multiphase nature of the outflowing material.

The dynamical evolution of a starburst-driven outflow has been studied using hydrodynamical simulations. The deposition of mechanical energy by supernovae and stellar winds creates an overpressured cavity of hot gas inside the starburst. This cavity expands, sweeping up ambient material and developing a bubble-like structure. If the ambient medium is stratified (like in a disk), the bubble expands most rapidly in the direction of the vertical pressure gradient. Numerical simulations show that when the bubble size reaches several disk vertical scale heights, it is fragmented because of Raleigh-Taylor instabilities which allow the hot gas to blow out of the disk into the halo in a weakly collimated bipolar outflow (a "wind"). ➋ *Figure 10-11* shows snapshots of the density distribution for models of galaxies with different initial masses and mechanical luminosities (as indicated in the legend). These simulations, carried out by Mac Low and Ferrara (1999), also showed that it is relatively difficult for supernovae-driven outflows to remove the gas from low-mass galaxies, while they appear to be fairly efficient at ejecting metals. This "metal loading" effect is just another manifestation of the importance of the multiphase nature of the outflow material and is often neglected in models of galaxy formation.

As mentioned above, supernovae feedback is believed to play a very important role in regulating the number of faint galaxies (Benson et al. 2003) but also in shaping the mass-metallicity relation that is observed for star-forming galaxies (Tremonti et al. 2004) and in enriching the intergalactic and intracluster medium (De Lucia et al. 2004b). Given the uncertainties discussed above, this process is included in galaxy formation models using a number of different prescriptions that are based on observations and/or theoretical arguments. Currently, it is difficult to argue that one specific model is and/or works better than another.

3.4.3 AGN Feedback

AGN can release huge amounts of energy during their lifetimes. Assuming an energy conversion efficiency of ϵc^2 per unit of accreted mass, one finds that an accreting black hole liberates $\sim 10^{19}(\epsilon/0.01)$ erg per gram, and it is easy to compute that this energy input can easily exceed the binding energy of the host galaxy (Begelman 2004). Broadly speaking, there are two different forms of feedback.

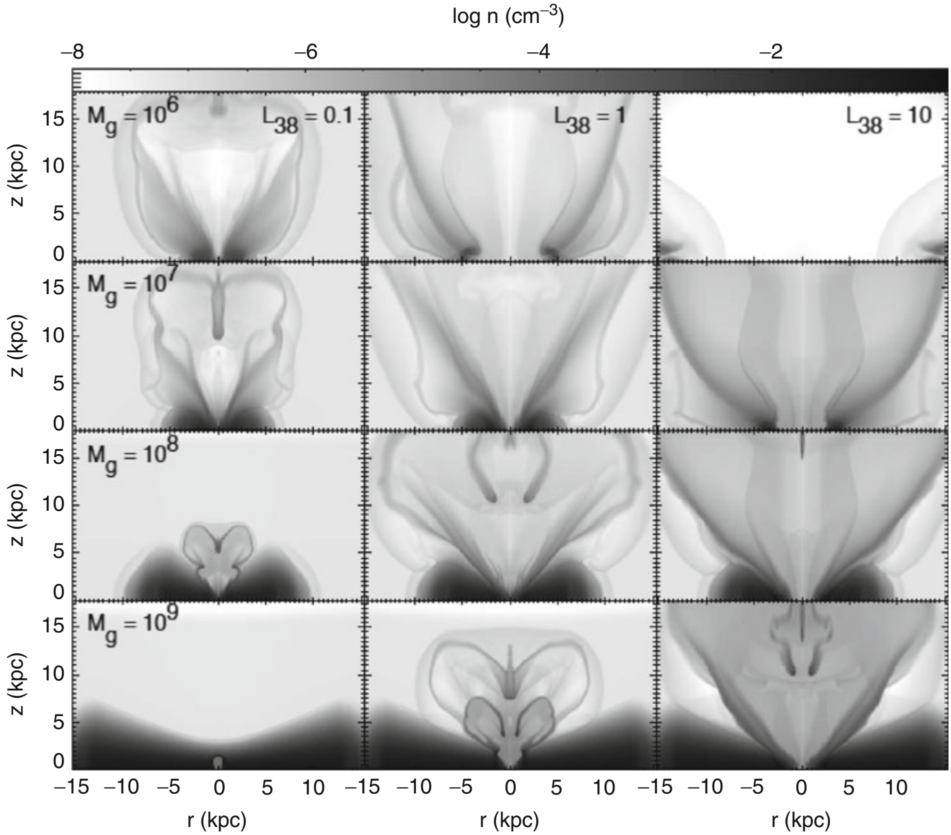


Fig. 10-11
From Mac Low and Ferrara (1999). Density distributions for models with different initial masses M_g and mechanical luminosities L in units of 10^{38} ergs s^{-1} after 75 Myr (energy input ended at 50 Myr)

The photons generated by AGN can ionize and heat the surrounding gas. If the host galaxy contains significant amounts of dust, the radiation pressure on the dust grains can overcome the gravitational force of the halo, generating a momentum-driven wind (Murray et al. 2005). This is a "radiative" feedback mode, and in the literature, it is sometimes referred to as "quasar mode." This energy injection mechanism is effective in broad-line quasars for which the presence of high velocity winds has been confirmed in a number of cases. Galaxy formation models incorporate this energy injection channel in mergers of gas-rich galaxies that can funnel copious amounts of cold gas toward the central regions of galaxies and feed the central black holes with high gas accretion rates. This particular form of feedback is therefore believed to be important at higher redshift where the activity of quasars peaks. Numerical simulations show that it can affect significantly the physical properties of the host galaxy, by expelling large amounts of gas and therefore suppressing significantly subsequent star formation (Springel et al. 2005a).

When the accretion rate onto the central black hole is low, AGN can drive highly collimated and powerful jets which can reach out well into the surrounding halo. This is a "mechanical" feedback mode, sometimes referred to as "radio mode." Evidence for this form of feedback can

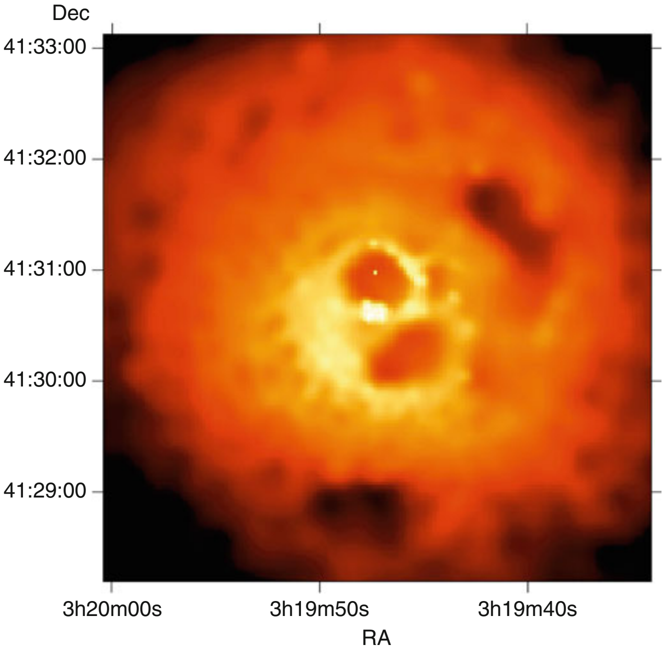


Fig. 10-12
From Fabian et al. (2000). Adaptively smoothed 0.5–7 keV Chandra image of the core of the Perseus cluster

be seen in the central regions of galaxy clusters: X-ray observations show that these often contain cavities that are filled with relativistic gas and are believed to be inflated by the jet launched from the central black hole. *Figure 10-12* shows, for example, an X-ray image (adaptively smoothed) of the central region of the Perseus cluster that contains a bright radio source at its center. It has long been realized that this form of feedback can provide a solution to the "cooling flow" problem, i.e., the observation that the gas at the center of most galaxy clusters is apparently not condensing and turning into stars, although the observed X-ray emission implies a cooling time that is much shorter than the age of the system (Tabor and Binney 1993). The ensemble-averaged power from radio galaxies seems sufficient to offset the mean level of cooling, and a large fraction of central cluster galaxies are radio loud (Best et al. 2007). The steep dependence of the radiative cooling function on density makes, however, difficult to stabilize cooling flows so that heating approximately balances cooling at all radii. Numerical simulations show that the efficiency of this feedback in suppressing gas condensation depends strongly on a number of unknown parameters, e.g., the duty cycle (i.e., the frequency of the energy injection), the geometry, and gas viscosity (e.g., Sijacki and Springel 2006).

In addition to the two modes of AGN feedback described above, significant outputs of energetic particles (cosmic rays) or exotic particles consisting, for example, of relativistic neutrons and neutrinos can contribute to inject energy into the surrounding gas. The precise composition of the bubbles, in these terms is, however, not well known, both from an observational and a theoretical point of view. Thus, it is currently unclear if the energy released via this channel is significant.

3.5 Chemical Enrichment

As explained above, the first generation of stars (the Pop III stars) formed from gas with primordial composition. Stellar nucleosynthesis and the subsequent pollution of the interstellar and intergalactic medium (through, for example, supernovae-driven winds) affect the formation of later stellar populations. In particular, the presence of heavy elements increases significantly the rate at which gas can cool and affects the luminosity and colors of the stellar populations. In addition, star formation also leads to the formation of dust which attenuates the optical and ultraviolet light of galaxies and re-emits at longer wavelengths.

The final stages of stellar evolution and metal production depend on the stellar mass. Stars with masses $\lesssim 8M_{\odot}$ end up their life as C/O white dwarfs, after an asymptotic giant branch (AGB) phase during which the star is burning helium in an inner shell and hydrogen in an outer shell. Unfortunately, there are still large theoretical uncertainties in the treatment of convection and mass loss from AGB stars. If the C/O white dwarf is part of a close binary system, it can accrete material from the companion. When the star reaches the Chandrasekhar limit ($1.4\,M_{\odot}$), it explodes as supernovae type Ia which dominate the production of elements in the iron peak. Massive stars (with masses $\gtrsim 8M_{\odot}$) enrich the interstellar medium with metals via both stellar winds and their final explosions as core-collapse supernovae (type II SNe). These are primarily responsible for the production of α elements (among which oxygen, magnesium, silicon, calcium) but also of other elements like nitrogen and sodium. Since the progenitors of type II SNe are massive stars with lifetimes shorter than $\sim 10^7$ year, while the progenitors of type Ia SNe are less massive stars with lifetimes $\gtrsim 10^8$ year, the relative proportions of the metal species they contribute (often quantified in the [α/Fe] ratio) provide information on the time scale of star formation. So studying the metallicity and [α/Fe] ratio of galaxies, it is possible to constrain the time scale over which star formation took place. The only complication is given by the fact that the [α/Fe] ratio does not depend only on the star formation history but also on the shape of the IMF.

In the framework of galaxy formation models, chemical evolution has often been (and still largely is) included using the instantaneous recycling approximation (i.e., the models neglect the finite lifetime of stars so that both chemical enrichment and gas recycling are assumed to take place at the same time of star formation) and a constant yield that is usually treated as a free parameter. More recent studies have included a more accurate treatment of type Ia supernovae and are able to follow the evolution of individual elements (e.g., Nagashima et al. 2005; Arrigoni et al. 2010).

3.6 Galaxy-Galaxy Interactions

In the hierarchical scenario, dark matter haloes (and therefore the galaxies that reside in them) undergo frequent interactions with each other. These interactions have dramatic influence on the morphologies and star formation histories of the galaxies involved. Numerical simulations have shown that close interactions can lead to a strong internal dynamical response driving the formation of spiral arms and, depending on the structural properties of the disks, of strong bar modes. The developing nonaxisymmetric structures (spiral arms and/or central bars) lead to a compression of the gas that can fuel starburst/AGN activity (see Mihos 2004, and references therein). Simulations have also shown that in sufficiently close encounters between galaxies of similar mass, violent relaxation completely destroys the disk and leaves a kinematically hot remnant with photometric and structural properties that resemble those of elliptical galaxies.

The merger hypothesis for the formation of elliptical galaxies was suggested early on by Toomre and Toomre (1972) and later confirmed by many numerical simulations (Mihos 2004; Cox et al. 2008, and references therein). In recent years, a large body of observational evidence has been collected that demonstrates that a relatively large fraction of early-type systems show clear evidence of interactions, mergers, and recent star formation, in particular at high redshift. However, the data also seem to indicate that only a small fraction of the final mass is involved in these episodes. This observational result has often been interpreted as strong evidence against the somewhat extended star formation history naively predicted from hierarchical models. A related issue concerns the α-element enhancements observed in elliptical galaxies. As explained above, the [α/Fe] ratio is believed to encode important information on the time scale of star formation, and it is a well-established result that massive ellipticals have supersolar [α/Fe] ratios, suggesting that they formed on relatively short time scales and/or have an initial mass function that is skewed toward massive stars. The inability of early models of the hierarchical merger paradigm to reproduce this observed trend has been pointed out as a serious problem for these models (Thomas 1999).

In order to model galaxy interactions and mergers, one needs to know what determines the structural and physical properties of a merger remnant. Numerical simulations have shown that these depend mainly on the following two factors:

1. The progenitor mass ratio. As mentioned above, during "major" mergers, violent relaxation plays an important role, and as a consequence, the merger remnant has little resemblance to its progenitors. On the other hand, during minor mergers, the interaction is less destructive so that the merger remnant often resembles its most massive progenitor. The exact value at which one distinguishes between minor and major mergers is somewhat arbitrary but is usually chosen to be of the order of $M_2/M_1 \sim 0.3$.
2. The physical properties of the progenitors. The structure of the galaxies involved in a merger plays an important role in determining the response to interactions: disks that are stable against the growth of instabilities (e.g., because of a central bulge or a lowered disk surface density) will be less "damaged" than disk-dominated systems that are prone to strong instabilities. In addition, in a merger between two gas-rich progenitors, a significant fraction of the gas content can be fuelled toward the center, triggering a starburst and/or accretion of gas onto the central black hole. Merger-driven starbursts are instead suppressed if the two merging systems are gas poor. These purely stellar mergers are often referred to as a "dry" or "red," and as will be discussed below, they are believed to contribute significantly to the recent assembly of elliptical galaxies.

➜ *Figures 10-13* and ➜ *10-14* show the projected stellar and gas mass density, respectively, during a merger with baryonic mass ratio 2.3:1. The figures show that the satellite galaxy first makes a fast, direct approach toward the primary galaxy. The tidal interaction between the two merging disks generates symmetric tails in both of them. Due to the initial orbital energy, the two galaxies separate again for several orbital periods (~1 Gyr in the particular case shown) before getting closer again. After the first or second passage, the initial angular momentum is lost, and the orbit becomes almost entirely radial. This limits the coupling between orbital and spin angular momentum and therefore the tidal response during the final coalescence.

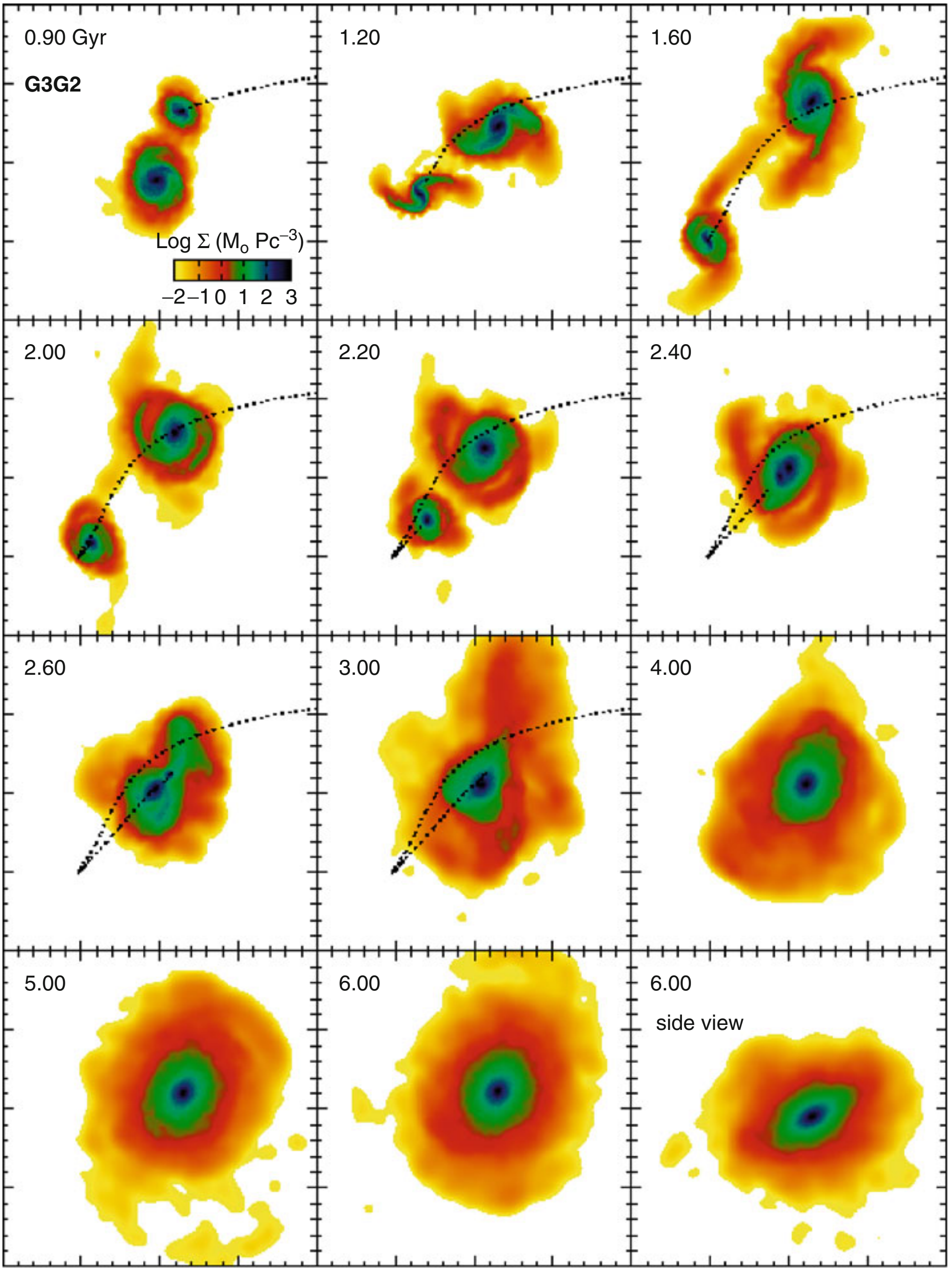


Fig. 10-13

From Cox et al. (2008). Projected stellar mass density during a merger simulation with mass ratio 2.3:1, as viewed in the orbital plane. Each panel measures 200 kpc on a side. The time (in Gyr) is displayed in the *upper left-hand* side of each panel. The orbit of the satellite galaxy G2 is denoted by a *dotted line* until it has completely merged with the primary galaxy. The *bottom right-hand panel* shows a side view of the final merger remnant

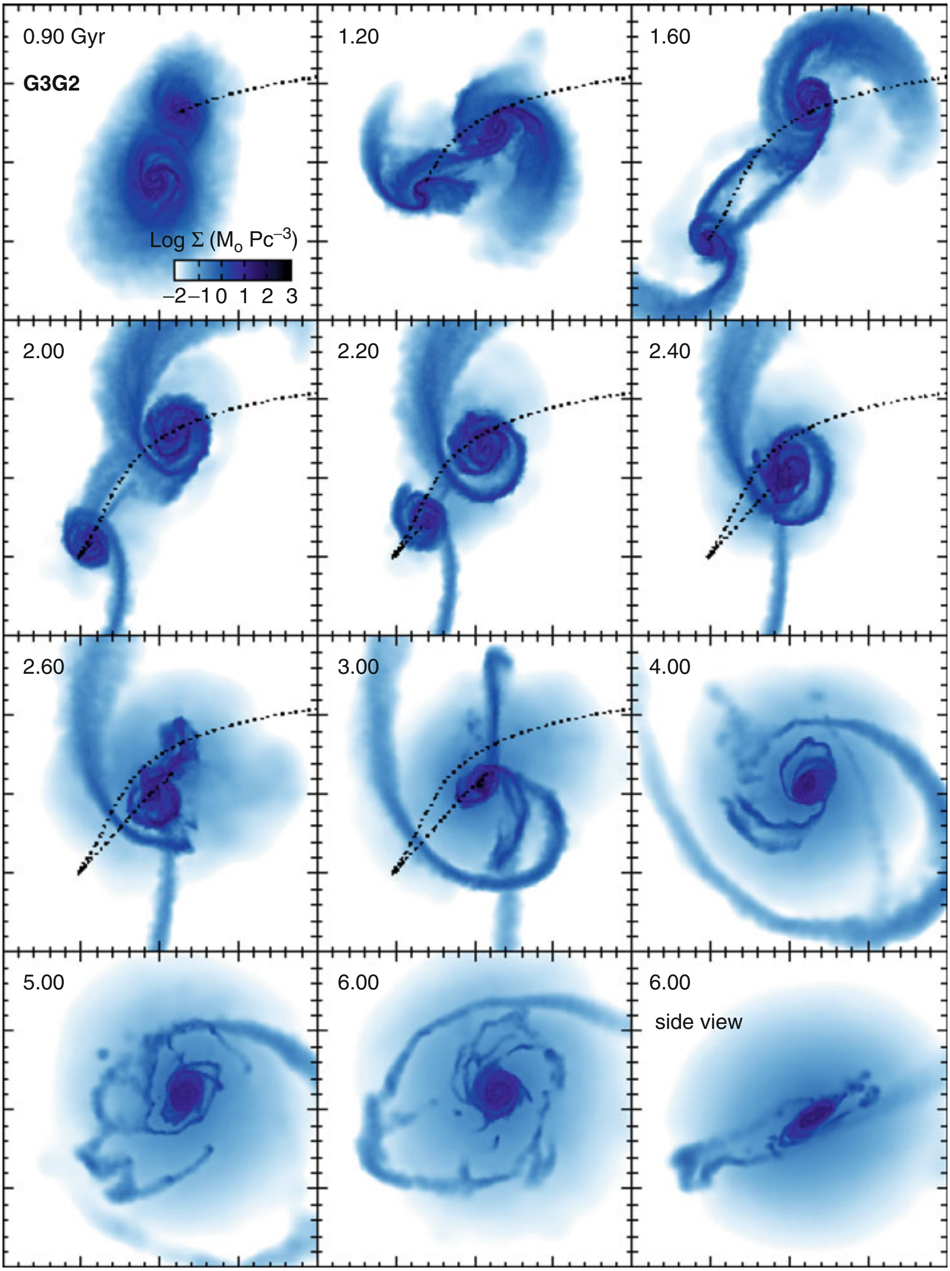


Fig. 10-14

From Cox et al. (2008). Similar to *Fig. 10-13* but for the gaseous component

The time scales of the galaxy mergers depend significantly on the orbital parameters that, to some extent, also affect the structural properties of the remnant. For example, the relative orientation of the orbital spin with respect to the intrinsic spin of the progenitors influences significantly the prominence of tidal tails. A good approximation of the merging times of galaxies is provided by the classical Chandrasekhar (1943) dynamical friction formula:

$$\frac{d}{dt}\vec{v}_{\mathrm{orb}} = -4\pi G^2 \ln(\Lambda)\, M_{\mathrm{sat}}\, \rho_{\mathrm{host}}(< v_{\mathrm{orb}})\, \frac{\vec{v}_{\mathrm{orb}}}{v^3_{\mathrm{orb}}}\,,$$

where $\rho_{\mathrm{host}}(<v_{\mathrm{orb}})$ is the density of background particles with velocities less than the orbital velocity v_{orb} of the satellite, M_{sat} is the mass of the satellite, and Λ is the Coulomb logarithm that depends on the mass ratio between the two merging galaxies. The formula given above, that is valid in the approximation of a point mass satellite and a uniform background mass distribution, is often adopted in analytic models of galaxy formation to estimate the time scale for an orbiting satellite to lose its energy and angular momentum and merge with the central galaxy of its host halo. Recent work has, however, shown that the classical dynamical friction formula tends to underestimate merging times computed from controlled numerical experiments and high-resolution cosmological simulations (e.g., Boylan-Kolchin et al. 2008). In addition, it should be noted that different models usually assume different variations of the classical formula (e.g., adopt a different "fudge" factor and/or a different expression for the Coulomb logarithm) that can lead to significant differences in the estimated merger times (see Sect. 8 and Fig. 14 in De Lucia et al. 2010).

3.7 The Environment

The distribution of galaxies on the sky is not uniform: galaxies appear to be arranged in a complex web of filaments and sheets that surround empty "voids" and intersect in dense nodes that can contain up to thousands galaxies, the rich clusters of galaxies. It has been known for a long time that the local and large-scale environments play an important role in determining many galaxy properties. First indications of a correlation between the galaxy type and the environment can be found in the *The Realm of Nebulae* by E. Hubble (1936), but the milestone paper in the subject is probably the work by Dressler (1980), who showed the existence of a well-defined relationship between local galaxy density and galaxy type for a sample of ~20 massive nearby clusters.

Disentangling the processes responsible for the observed correlations has proved difficult, and it remains unclear whether the observed relations are imprinted during formation or by physical processes at work preferentially in dense environments. The difficulty is in part intrinsic: according to the current paradigm for structure formation, dark matter collapses into haloes in a bottom-up fashion. Small systems form first and subsequently merge to form progressively larger systems. As structure grows, galaxies join more and more massive systems, therefore experiencing different "environments" during their lifetimes. In this context, it is clear that both "heredity" (i.e., the initial conditions) and "environment" (i.e., subsequent physical processes that galaxies experience during their lifetimes) do play a role in shaping the observed galaxy properties and in determining the observed environmental trends.

A number of different physical mechanisms have been early identified that can influence significantly the physical properties of galaxies in a cluster environment. Broadly speaking, they can be grouped in two big families: (i) interactions with other cluster members and with the cluster potential well, and (ii) interactions with the hot gas that is known to permeate galaxy clusters. In the following, the specific mechanisms often considered when trying to assess the influence of the environment on galaxy evolution are discussed in more detail.

3.7.1 Galaxy Harassment

Galaxy mergers and more generally strong galaxy–galaxy interactions are commonly viewed as a rarity in massive clusters because of the large velocity dispersion of the system. It should be noted, however, that they are still important in the outskirts of galaxy clusters, and they were certainly more efficient in the infalling group environment. Therefore, they may represent an important "preprocessing" step in the evolution of cluster galaxies. In rich clusters, the encounters between galaxies will be generally high-speed interactions. The colliding galaxy is impulsively heated and becomes less bound and more vulnerable to disruptions by further encounters and by tidal interactions with the global cluster potential.

The cumulative effect of repeated, numerous fast encounters is usually referred to as "harassment." This process has been discussed in early work on the dynamical evolution of cluster galaxies (Richstone 1976) and has been explored in detail using numerical simulations (Farouki and Shapiro 1981; Moore et al. 1998). These have confirmed that repeated high-speed collisions, coupled with the effects of the global tidal field of the cluster, can drive a strong response in cluster galaxies. The efficiency of the process is, however, largely limited to low-luminosity hosts due to their slowly rising rotation curves and their low-density cores. Therefore, it is believed that harassment might have an important role in the formation of dwarf ellipticals or in the destruction of low-surface brightness galaxies in clusters, but it is less able to explain the evolution of luminous cluster galaxies.

➲ *Figure 10-15* shows the evolution of the stellar surface density of a galaxy that is orbiting close to the center of a galaxy cluster. The first stages of the evolution are characterized by the formation of a strong bar and of an open spiral pattern that is, however, easily stripped by tidal interactions. In contrast, the bar appears to be quite stable. It undergoes strong "buckling" instabilities that make the central part of the galaxy more spherical. In the final stages, the galaxy resembles a drawf spheroidal system.

3.7.2 Cannibalism

Early theoretical studies have discussed the role of cannibalism due to dynamical friction in the formation of brightest cluster galaxies (e.g., Ostriker and Tremaine 1975). This early work, however, significantly overestimated the efficiency of the process due to different simplified assumptions adopted. In the now standard paradigm of structure formation, clusters assemble quite late, through the merging of smaller systems. In this perspective, cooling flows are the main fuel for galaxy formation at high redshift, in dense and lower-mass haloes. This source is removed at lower redshift possibly due to feedback from AGN. Galaxy-galaxy mergers, as discussed above, are most efficient within small haloes with low-velocity dispersions. These mergers are indeed driven by dynamical friction, but it is the accretion rate of the galaxies onto the protocluster, along with the cluster growth itself, that regulate and set the conditions for galaxy merging.

This is illustrated very nicely in ➲ *Fig. 10-16* which shows the merger tree of the central galaxy of a cluster-sized halo (for details, see De Lucia and Blaizot 2007). The brightest cluster

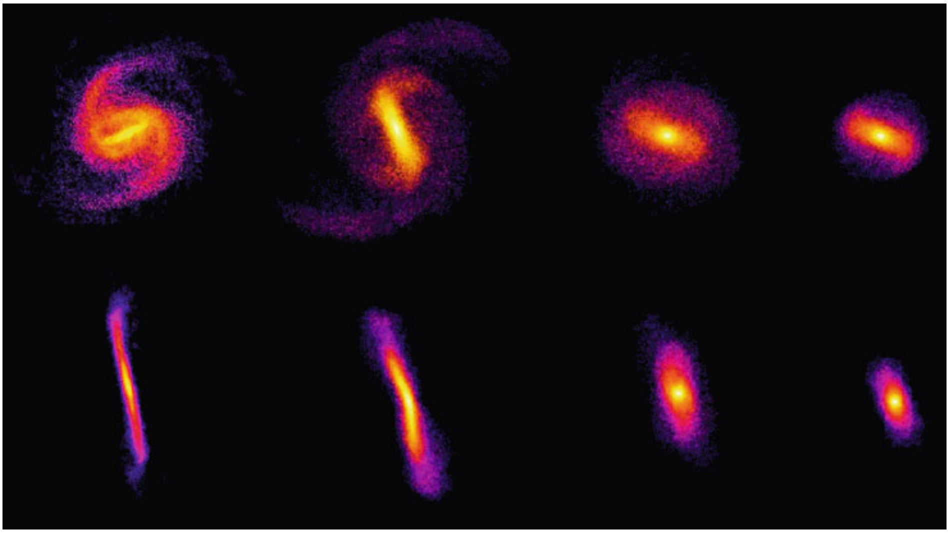


Fig. 10-15

From Mastropietro et al. (2005). Evolution of the stellar surface density of a galaxy that is orbiting close to the center of a galaxy cluster at $z = 0$. The *top panels* represent the face on projections, while the edge on projections are shown in the *bottom panels*

galaxy (BCG) itself lies at the top of the plot (at $z = 0$), and all its progenitors (and their histories) are plotted downward going back in time recursively. Galaxies with stellar mass larger (resp. smaller) than $10^{10}\ h^{-1}\ \mathrm{M}_{\odot}$ are shown as symbols (resp. lines) and are color-coded according to their rest-frame B-V color. The leftmost branch in ➋ *Fig. 10-16* represents the "main branch," obtained by connecting the galaxy at each time step to the progenitor with the largest stellar mass at the immediately preceding time step (the "main progenitor").

➋ *Figure 10-16* shows another important point: in the context of the hierarchical paradigm for structure formation, the full history of a galaxy is described by its complete merger tree. Whereas in the "monolithic" approximation, the history of a galaxy can be described by a set of functions of time, hierarchical histories are difficult to summarize in a simple form because even the identity of a galaxy is ill-defined. A galaxy is no more a single object when viewed at different times but the ensemble of its progenitors, all of which need to be taken into account for a correct characterization of the stellar population of the final object. It is also interesting to note that although the merger trees of these central galaxies have a very large number of branches, only a small fraction of these contribute significantly to the buildup of their stellar mass: in this particular example, ~70% of the mass comes from the accretion of 12 galaxies more massive than $10^{10}\ h^{-1}\ \mathrm{M}_{\odot}$ (see also De Lucia et al. 2006).

3.7.3 Ram-Pressure Stripping

Galaxies travelling through a dense intracluster medium suffer a strong ram-pressure stripping that can sweep cold gas out of the stellar disk (Gunn and Gott 1972). Depending on the binding energy of the gas in the galaxy, the intracluster medium will either blow through the galaxy, removing some of the diffuse interstellar medium, or will be forced to flow around the galaxy

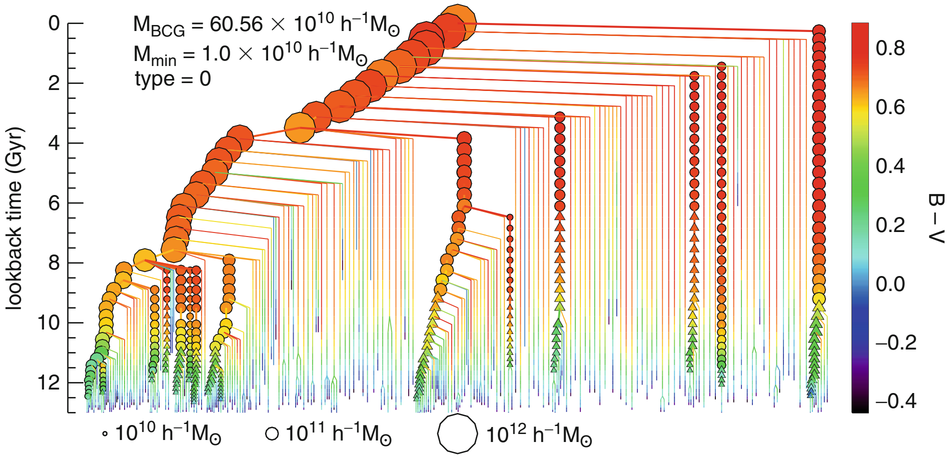


Fig. 10-16

From De Lucia and Blaizot (2007). BCG merger tree. Symbols are color-coded as a function of B-V color and their area scales with the stellar mass. Only progenitors more massive than 10^{10} $M_{\odot}$ h^{-1} are shown with *symbols*. *Circles* are used for galaxies that reside in the FOF group inhabited by the main branch. *Triangles* show galaxies that have not yet joined this FOF group

(Cowie and Songaila 1977; Nulsen 1982). Ram-pressure stripping is expected to be more important at the center of massive systems because of the large relative velocities and higher densities of the intracluster medium. By considering the distribution and history of ram-pressure experienced by galaxies in clusters, Brüggen and De Lucia (2008) estimated that strong episodes of ram-pressure are indeed predominant in the inner core of the clusters. They also showed, however, that virtually all cluster galaxies suffered weaker episodes of ram-pressure, suggesting that this physical process might have a significant role in shaping the observed properties of the entire cluster galaxy population. In addition, Brüggen and De Lucia found that ram-pressure fluctuates strongly so that episodes of strong ram-pressure alternate to episodes of weaker ram pressure, possibly allowing the gas reservoir to be replenished and intermittent episodes of star formation to occur.

Ample observational evidence that ram-pressure is occurring is available, and the process has been extensively studied using hydrodynamical simulations. ➋ *Figure 10-17* shows snapshots from the simulations carried out by Quilis et al. (2000). The figure shows a galaxy moving face on and nearly edge on through the core of a rich cluster at a velocity of ~2,000 km s^{-1}. These simulations showed that the time scale for stripping is very short compared to the orbital time scale, and that the multiphase structure of the interstellar medium and the presence of bubbles and holes make the disk more susceptible to viscous stripping. A simple estimate of the efficiency of ram-pressure was obtained by Gunn and Gott (1972), by comparing the ram pressure with the galactic gravitational restoring force per unit area. This leads to the following condition:

$$\rho_{\rm ICM} > \frac{2\pi G \Sigma_{\star} \Sigma_{\rm ISM}}{V^2}$$

where $\rho_{\rm ICM}$ is the density of the intracluster medium, V is the velocity of the galaxy, and $\Sigma_{\rm star}$ and $\Sigma_{\rm ISM}$ are the mean stellar and gaseous surface density of the disk. Early numerical

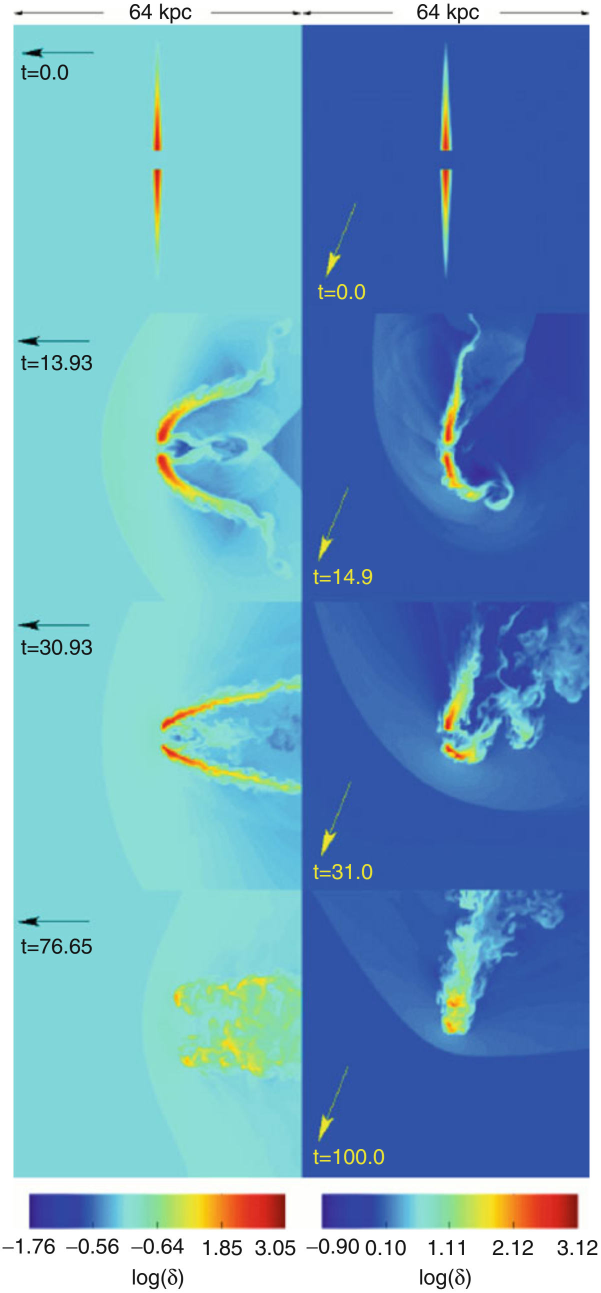


Fig. 10-17

From Quilis et al. (2000). The evolution of the gaseous disk of a spiral galaxy moving face on (*left column*) and inclined 20° to the direction of motion (*right column*) through a diffuse hot intracluster medium. Each *snapshot* shows the density of gas ($\delta = \rho/\rho_{ICM}$) within a 0.2-kpc slice through the center of the galaxy, and each frame is 64 kpc on a side

simulations showed that this analytical estimate fares fairly well, as long as the galaxies are not moving close to edge on. More recent numerical work (e.g., Roediger and Brüggen 2007) has shown that this formulation often yields incorrect mass-loss rates. In addition, simple models based on the Gunn and Gott formula usually do not consider the possibility that ram-pressure stripping could temporarily enhance star formation.

3.7.4 Strangulation

Current theories of galaxy formation assume that when a galaxy is accreted onto a larger structure, the gas supply can no longer be replenished by cooling that is suppressed because of the removal (by tides and ram-pressure) of the hot-gas halo associated with the infalling galaxy (Larson et al. 1980). This process is usually referred to as "strangulation" (or "starvation" or "suffocation"). It is common to read in discussions related to these physical mechanisms that strangulation is expected to affect the star formation of cluster galaxies on relatively long time scales, and therefore to cause a slow decline of the star formation activity. As we will see below, however, in recent galaxy formation models, this process is usually associated to a strong supernovae feedback. As a consequence, galaxies that fall onto a larger system consume their cold gas rapidly, moving onto the red sequence on very short time scales.

Traditionally, in galaxy formation models, the stripping of the hot gaseous reservoir has been assumed to be complete and instantaneous. Using a suite of controlled full hydrodynamic simulations, however, McCarthy et al. (2008) have found that a fraction (about 30%) of the initial hot galactic halo gas can remain in place even after 10 Gyr. Saro et al. (2010) have confirmed that cooling can occur on satellite galaxies, but this seems to be limited to the most massive ones. In these satellites, the star formation can last for up to ~1 Gyr after accretion, albeit significantly suppressed with respect to the average value before accretion.

3.8 Stellar Populations

Observational studies of galaxies make use of the radiation emitted by them to infer their physical properties. In order to make a close link between model predictions and observational data, it is therefore necessary to compute the luminosity emitted by the galaxy as a function of wavelength or frequency. Analytically, the spectral energy distribution of a galaxy can be expressed as the superposition of numerous "single-stellar populations" (SSPs) that are populations of stars with the same age, initial mass function, and chemical composition. The luminosity of each of these SSPs can be written as

$$L_\nu^{(\mathrm{SSP})}(t, Z, \phi) = \int_{M_{\mathrm{min}}}^{M_{\mathrm{max}}} \phi(M') L_\nu^{(\mathrm{star})}(t, Z) dM'$$

where M_{min} and M_{max} are the minimum and maximum mass for stars, respectively, $\phi(M)$ is the initial mass function, and $L_\nu^{(\mathrm{star})}(t, Z)$ is the spectrum of a single star of metallicity Z and age t. If the luminosity of the SSPs is known as a function of age and metallicity, then the luminosity of a galaxy can be written as

$$\begin{aligned} L_\nu^{(\mathrm{galaxy})} &= \int_0^t dt' \int_0^\infty dZ' \dot{M}_\star(t', Z') \\ &\quad \times L_\nu^{(\mathrm{SSP})}(t - t', Z', \phi) \end{aligned} \tag{10.5}$$

where $\dot{M}_\star(t, Z)$ gives the rate at which stars of metallicity Z form at the time t inside the galaxy.

Several libraries are available in the literature which provide $L_\nu^{(\mathrm{SSP})}(t, Z, \phi)$ for different ages, metallicities, and initial mass functions (e.g., Bruzual and Charlot 2003; Maraston et al. 2009; Conroy et al. 2009). These libraries are constructed using a combination of theoretical stellar evolution models, direct observations of stars for which age and metallicity can be measured, and theoretical models of stellar atmospheres where no observations are available. In the framework of galaxy formation, these population synthesis models are usually treated as "black boxes." It is important, however, to remember that significant uncertainties remain in many of their ingredients. The AGB regime, for example, is very difficult to treat because of the pulsational regime, the double-shell burning, and especially the strong mass losses affecting this phase. This leads to large uncertainties in the evolution of the spectral continuum in rest-frame near infrared (for a review, see Maraston 2011).

Real galaxies are not made only of stars but also contain gas (both hot and cold) and dust. This can significantly affect the observed luminosities in the ultraviolet and in the optical and even dominate the luminosity in the far-infrared portion of the spectrum. Indeed dust, that is believed to be produced in the envelops of AGB stars and from supernovae, absorbs light emitted by stars particularly at short wavelengths, is heated by this light, and re-emits it at longer wavelengths (infrared and sub-mm). It is clear that in order to accurately model the dust extinction and emission, one needs to know how dust grains and stars are distributed and which is the composition of the dust grains.

For a population of galaxies that are assumed to have the same composition and distribution of dust, one can derive an "effective" extinction law that can then be used without modeling in detail the dust distribution. For example, one can assume that an "obscuting screen" or a "slab" geometry of dust is sitting between the galaxy and the observer. A simple estimate of the amount of extinction can then be obtained by adopting the measured effective law (e.g., the law found for local starburst galaxies by Calzetti et al. 1994) and by scaling the depth at optical wavelengths on the basis of the physical properties of the galaxy under consideration (e.g., gas content and metallicity). Alternatively, the propagation of light through the interstellar medium can be studied using radiative transfer calculations which take into account the geometry of the galaxy, as well as the distribution and mix of dust (Silva et al. 1998; Jonsson 2006). Recent work by Fontanot et al. (2009b) has compared the two approaches and has shown that the former can predict quite accurately results from the full radiative transfer calculation, with a small scatter. However, there is a large galaxy-to-galaxy variation, likely due to different geometries, that the simple approach cannot capture. It should be noted that also radiative transfer codes often need to make a number of assumptions about the physical state of the dust and about its distribution relative to stars and the interstellar medium. Finally, given the uncertainties involved, often these properties are assumed not to vary as a function of cosmic time.

4 Putting It Together: Models of Galaxy Formation in a Cosmological Context

As discussed above, the process of galaxy formation involves complex and nonlinear physical processes that cover many orders of magnitude in physical size (from the scale of black holes to that of massive galaxy clusters) and in time scales. In addition, as we have seen in the previous section, many (if not all) of the physical processes at play cannot be treated from first principles.

In the past decades, however, three major approaches have been used and further developed in order to circumvent these difficulties and improve our understanding of the physical processes that drive galaxy formation and evolution. The following provides a brief review of these techniques and discusses the most recent successes and problems of one particular class of models that is widely used to make detailed predictions of galaxy properties at different cosmic epochs and environments.

This section will not provide a detailed description of the implementations used in different models. Given the complexity involved, such a description would rapidly become out of date as models are continuously being improved and developed. Rather, this section is aimed at discussing the weaknesses and the strengths of each of the methods that can be used to model galaxy formation in a cosmological context.

4.1 The Halo Occupation Distribution Method

This method essentially bypasses any explicit modeling of the physical processes driving galaxy formation and evolution and specifies the link between dark matter haloes and galaxies in a purely statistical fashion. The halo occupation framework has a long history: a first description can be found in Neyman and Scott (1952) who discussed an analytic model that described galaxy clustering as the superposition of randomly distributed clusters with given profiles. The method has become very popular in more recent years, after it was realized that it provides a powerful formalism for understanding the clustering of galaxies (e.g., Benson et al. 2000; Berlind and Weinberg 2002, and references therein). A classical halo occupation distribution (HOD) model can be constructed by specifying the probability that a halo of mass M contains N galaxies of a particular class (the halo occupation distribution – $P(N|M)$) and by assuming a spatial distribution of galaxies inside dark matter haloes (the most common assumption is that the distribution of galaxies follows that of the dark matter). The halo occupation distribution can then be constrained using galaxy clustering data. For example, a simple model that is often employed in the literature assumes that the mean number of galaxies above a certain luminosity threshold changes with halo mass as

$$N_{\rm avg} = \begin{cases} 0 & \text{if } \; M < M_{\rm min} \\ (M/M_1)^{\alpha} & \text{otherwise,} \end{cases} \tag{10.6}$$

where $M_{\rm min}$ is a cutoff halo mass below which haloes cannot contain galaxies. In this model, M_1 corresponds to the mass of haloes that contain, on average, one galaxy. ➲ *Figure 10-18* shows the influence of $M_{\rm min}$ (left panel) and α (right panel) on the galaxy correlation function and demonstrates that these observational data can be used to constrain the HOD parameters.

The same approach can be extended to constrain the halo occupation as a function of some galaxy physical property (e.g., luminosity, color, type, etc.). For example, one can define a “conditional luminosity function” $\Phi(L|M)dL$ that specifies the average number of galaxies with luminosities in the range $L \pm dL/2$ that reside in a halo of mass M. This provides a direct link between the observed galaxy luminosity function and the halo mass function:

$$\Phi(L) = \int_0^{\infty} \Phi(L|M)n(M)dM$$

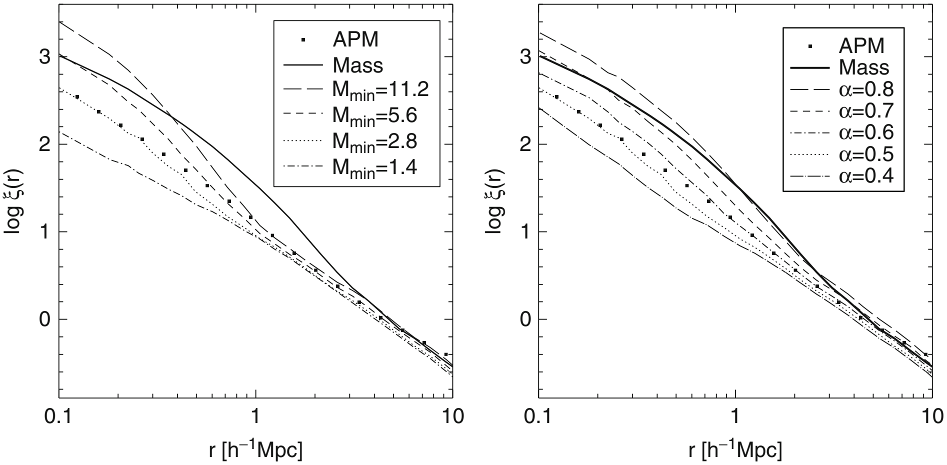


Fig. 10-18

From Berlind and Weinberg (2002). Influence of $M_{\min}$ and α on the predicted galaxy correlation function. *Curves* show galaxy correlation functions for HOD models constructed assuming the distribution described in (10.6) with different values of $M_{\min}$ and α as indicated in the legend. *Data points* show the correlation function measured from the APM galaxy survey (Baugh 1996)

In addition, one can express the total luminosity of a halo of a given mass as a function of the conditional luminosity function:

$$< L(M) > = \int_0^\infty \Phi(L|M) L dL$$

It has been shown that by adopting this formalism, it is possible to constrain both galaxy formation and cosmology by using the following observational data: the observed luminosity function, the luminosity dependence of the galaxy-galaxy two-point correlation function, and the average mass-to-light ratios as function of halo mass (van den Bosch et al. 2003).

The method described above is conceptually simple and relatively easy to implement. As shown, it can be constrained using the increasing amount of available information on the clustering properties of galaxies at different cosmic epochs, and it provides important statistical constraints for galaxy formation models. It remains difficult, however, to move from this purely statistical characterization of the link between dark matter haloes and galaxies to a more physical understanding of the galaxy formation process itself. In addition, the method described above implicitly assumes that the number of galaxies of a given type populating a dark matter halo, as well as the clustering properties of dark matter haloes, depend only on the halo mass. That is, the method neglects the assembly bias that has been discussed in Sect. 2.4. Recent studies show that this might not be a significant problem, at least for relatively bright galaxies (Tinker et al. 2008, and references therein). Further investigations are, however, needed, particularly in light of the statistical power and redshift coverage of forthcoming observational surveys.

A variant of the HOD approach is provided by the subhalo abundance matching (SHAM) technique. The method consists in assigning observable galaxy properties to the subhalo population of an N-body simulation, assuming a monotonic relation between these properties and some property of the substructure (e.g., the mass or the maximum circular velocity of dark matter halos) at the time of "accretion," i.e., when the halo was accreted onto a larger structure becoming a subhalo (Conroy et al. 2006; Wang et al. 2006). This method offers some advantages with respect to the simple HOD approach described above. For example, it explicitly accounts for the dependence of the halo history on the environment. It could, however, depending on the resolution of the simulation, miss a significant fraction of the galaxy population: the "orphan" galaxies (i.e., those whose parent substructures were destroyed below the resolution by tidal stripping).

4.2 Hydrodynamical Simulations

Two different approaches can be used to include gas physics in N-body simulations. The most straightforward technique is adopted in smoothed-particle hydrodynamics (SPH) codes. These are based on a Lagrangian method which essentially works dividing the fluid into a set of discrete elements (particles). These have a spatial distance ("smoothing length"), over which their properties are smoothed by a kernel function. Any physical quantity of a particle (for example, density, temperature, and chemical composition) can then be obtained by summing the relevant properties of all the particles which lie within the range of the kernel. The contributions of each particle to a physical property are weighted according to their distance from the particle of interest and their density. Because of the smoothing, SPH codes have problems in resolving and treating dynamical instabilities developing at sharp interfaces in a multiphase fluid (e.g., shocks, Agertz et al. 2007). The Lagrangian nature of the method, however, means that regions of high density are automatically better resolved than regions of low density so that it is possible to study many orders of magnitude in the fluid properties. An alternative scheme is adopted in Eulerian codes in which the fluid equations are solved on a grid which is fixed in time and that can be "refined" several times to increase the resolution in regions of interest. This method is thus well adapted for capturing shocks and discontinuities. The resolution of the simulation can be increased by using "adaptive mesh refinements," but it can become quite time consuming.

As a tool for studying galaxy formation and evolution, hydrodynamical simulations offer the great advantage of providing an explicit description of the gas dynamics. They are, however, quite demanding in terms of computational time and memory consumption so that it is often necessary to limit the resolution range and the size of the volume being simulated. Additionally, and perhaps more importantly, complex physical processes such as star formation, feedback, etc., still have to be included as "subgrid physics." This is the case either because the resolution of the simulation is inadequate to treat a specific problem or simply because (and this is true almost always) we do not have a complete theory for the physical problem under consideration.

Current state-of-the-art full hydrodynamic cosmological simulations include the Galaxies-Intergalactic Medium Interaction Calculation project (GIMIC, Crain et al. see 2009) and the OverWhelmingly Large Simulations project (OWLS, Schaye et al. 2010). In the GIMIC project, five different regions with different mean overdensities have been selected from the Millennium simulation. These have been resimulated at high resolution using a SPH code that takes into account metal-dependent cooling in the presence of an ionizing UV background and includes a model for star formation and supernova feedback. Since each of the simulations is a considerable investment of computational time, they have been run using a unique set of parameters and concentrating on the environmental effects of the physical processes considered. The OWLS project represents a complimentary approach as it is based on a suite of over 50 cosmological

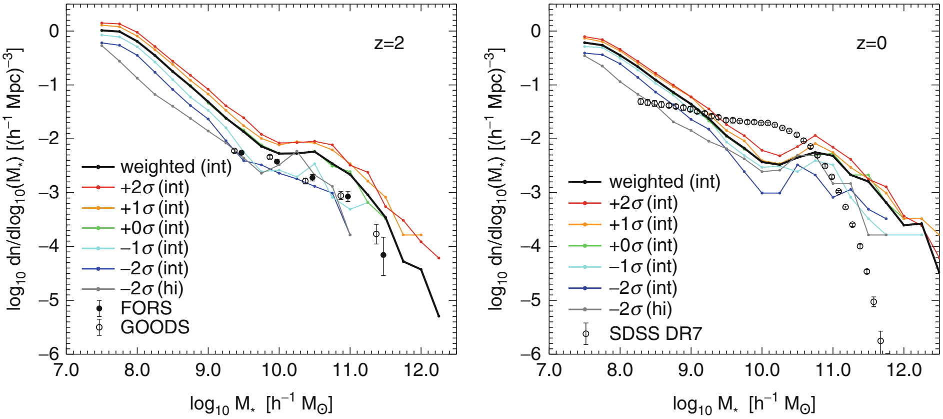


Fig. 10-19

From Crain et al. (2009). The stellar mass function of galaxies at $z = 2$ (*left panel*) and $z = 0$ (*right panel*). Results are shown for all five intermediate-resolution simulations considered in the GIMIC project (*colored curves*) and their weighted average (*black curve*). The stellar mass function of the region with largest overdensity at high resolution is also shown (*gray curve*) to illustrate the degree of convergence. Symbols with error bars show observational measurements from Drory et al. (2005) at $z = 2$ and from Li and White (2009) at $z = 0$

simulations (typically much smaller than the GIMIC high-resolution regions) that investigate the effects of different implementations of subgrid physics.

Figure 10-19 shows how the stellar mass function resulting from the GIMIC simulations compares with observational measurements at $z \sim 2$ in the left panel and at $z = 0$ in the right panel. Colored lines show results from each simulation while the black lines show their weighted average. The figure shows that the shape of the predicted stellar mass function differs significantly from that measured: the simulations predict an excess of low- and high-mass galaxies with respect to the observations and a “dip” in correspondence of the “knee” of the observed stellar mass function (that is where most of the galaxy mass is). At higher redshift, where the observations span only a limited mass range, the agreement looks better, but the shape of the predicted galaxy mass function does not vary with respect to the $z = 0$ predictions. The excess at large masses in the overdense regions originates mainly from the fact that these simulations do not include any modeling of the heating processes that can quench cooling flows in clusters (see also next sections). At low and intermediate masses, the disagreement with observational data is likely due to the simple wind model that has been adopted (for details, see Crain et al. 2009). A feedback model that follows the scalings expected for momentum-driven winds can give a better match with observational data around the knee of the luminosity function but still fails at higher and lower masses (Davé et al. 2011).

Much work has been done using direct simulations of the baryonic physics to study the formation and evolution of individual haloes at high resolution. These simulations take advantage of the zoom technique: first, a cosmological simulation of a large region is used to select a suitable target halo. The particles in the selected halo and its surroundings (usually all the particles within two times the virial radius) are then traced back to their initial Lagrangian region and are replaced by a larger number of lower mass particles. These are perturbed using the same fluctuation distribution as in the parent simulation but now extended to smaller scales to account for the increase in resolution. This resampling of the initial conditions thus allows a localized increase in mass and force resolution. Outside the high-resolution region, particles of variable mass (increasing with distance) are used, so that the computational effort is concentrated on the region of interest while still maintaining a faithful representation of the large-scale density and velocity field of the parent simulation.

On the cluster scale, significant disagreements with the observational data are still found in terms of the statistical description of the cluster galaxy population. For example, Saro et al. (2006) analyze a set of 19 cluster resimulations carried out using a SPH code that includes gas cooling, star formation, a detailed treatment of stellar evolution and chemical enrichment, as well as supernova feedback. They find that the total number of galaxies in their simulated clusters falls short of the observational measurements by a factor 2–3. The problem does not have an obvious numerical origin (e.g., lack of mass and force resolution). In addition, the BCGs of the simulated clusters are always predicted to be too massive and too blue when compared to data, stressing the need for the inclusion of a physical process that suppresses gas condensation at the center of relatively massive haloes.

At galaxy scales, simulations have generally had problems reproducing disk-dominated galaxies in typical dark matter haloes, when taking into account the cosmological setting. One major problem is known as the "angular momentum catastrophe": baryons condense early in clumps that then fall into larger haloes and merge via dynamical friction. This produces a net and significant transfer of angular momentum from the baryons to the dark matter. As a result, simulated disks are generally too small with up to ten times less angular momentum than real disk galaxies. The formation of a realistic rotationally supported disk galaxy in a fully cosmological simulation is still an open problem. Recent numerical work shows that it is in part due to limited resolution and related numerical effects that cause artificial angular momentum loss and spurious bulge formation (for a detailed discussion, see Mayer et al. 2008). The physics of galaxy formation during the merger of the most massive protogalactic lumps at high redshift and, in particular, the feedback due to supernovae is, howevelr, also playing a very important role (e.g., Scannapieco et al. 2008, and references therein).

4.3 Semianalytic Models of Galaxy Formation

The backbone of any semianalytic model is a statistical representation of the growth of dark matter haloes, i.e., a merger tree. Once the backbone of the model is constructed, galaxy formation and evolution can be coupled to the merger trees using a set of analytic laws that are based on theoretical and/or observational arguments to describe complex physical processes like star formation, supernovae and AGN feedback processes, etc. Adopting this formalism, it is possible to express the full galaxy formation process through a set of differential equations that describe the variation in mass of the different galactic components (e.g., gas, stars, metals). Given our limited understanding of the physical processes at play, these equations contain "free parameters," whose values are typically chosen in order to provide a reasonably good agreement with observational data in the local universe. These techniques find their seeds in the pioneering work by White and Rees (1978), have been laid out in a more detailed form in the early 1990s (White and Frenk 1991; Cole 1991), and have been substantially extended and refined in the last years by a number of different groups. For a detailed review of these techniques, the interested reader is referred to Baugh (2006).

In their first renditions, semianalytic models relied on Monte Carlo realizations of merging histories of individual objects, generated using the extended Press-Schechter theory (e.g., Kauffmann et al. 1993). An important advance of later years came from the coupling of semianalytic techniques with large-resolution N-body simulations that are used to specify the location and evolution of dark matter haloes – the birthplaces of luminous galaxies (Kauffmann et al. 1999; Benson et al. 2000). On a next level of complexity, some more recent implementations of these techniques have explicitly taken into account dark matter substructures, i.e., the haloes within which galaxies form are still followed when they are accreted onto a more massive system (Springel et al. 2001; De Lucia et al. 2004b). There is one important caveat to bear in mind regarding these methods: dark matter substructures are fragile systems that are rapidly and efficiently destroyed below the resolution limit of the simulation (see ➋ Sect. 2.4). Depending on the resolution of the simulations used, this can happen well before the actual merger can take place. This treatment introduces a complication due to the presence of "orphan galaxies," i.e., galaxies whose parent substructure mass has been reduced below the resolution limit of the simulation. In most of the available semianalytic models, these galaxies are assumed to merge onto the corresponding central galaxies after a residual merging time which is given by some variation of the classical dynamical friction formula.

One great advantage of these hybrid methods, with respect to classical techniques based on the extended Press-Schechter formalism, is that they provide full dynamical information about model galaxies. Using realistic mock catalogs generated with these methods, accurate and straightforward comparisons with observational data can be carried out. Since N-body simulations can handle large numbers of particles, the hybrid approach can access a very large dynamic range of mass and spatial resolution, at small computational costs. In addition, since the computational times are limited, these methods also allow a fast exploration of the parameter space and an efficient investigation of the influence of specific physical assumptions. This comes at the expenses, however, of loosing an explicit description of the gas dynamics.

One common criticism to semianalytic models is that there are "too many" free parameters. It should be noted, however, that the number of these parameters is not larger than the number of published comparisons with different and independent sets of observational data, for any of the semianalytic models discussed in the recent literature. In addition, these are not "statistical" parameters but, as explained above, they are due to our lack of understanding of the physical processes considered. Therefore, a change in any of these parameters has consequences on a number of different predictable properties so that often there is little parameter degeneracy for a given set of prescriptions. Finally, observations and theoretical arguments often provide important constraints on the range of values that different parameters can assume.

4.4 Successes and Problems of Semianalytic Models of Galaxy Formation

Clearly, each of the methods described above has its own advantages and weaknesses, and they should be viewed as complementary rather than competitive. In the framework of galaxy formation, semianalytic models certainly represent the most developed theoretical tool for interpreting observations of galaxy formation and evolution. This section provides a brief discussion of some of the most recent successes and problems of current models.

It is interesting to start this discussion from what can be considered the most fundamental description of the galaxy population: the galaxy luminosity function. As mentioned above, since early implementations of semianalytic techniques, it was clear that a relatively strong supernovae feedback was needed in order to suppress the large excess of faint galaxies due to the steep increase of low-mass dark matter haloes (White and Frenk 1991; Benson et al. 2003). The left panel of *Fig. 10-20* shows results from different models: the simplest one is obtained converting the dark matter halo mass function into a galaxy luminosity function by assuming a fixed mass-to-light ratio (this is the same model shown in *Fig. 10-9*). As discussed in Sect. 3.4, this model overpredicts the faint and the bright ends of the luminosity function by orders of magnitude. The other lines shown in the left panel of *Fig. 10-20* correspond to different models where different ingredients have been switched on, as indicated in the legend. None of these models reproduces the observational measurements. The right panel of the figure shows how the predicted K-band luminosity function compares with observational measurements, for increasing efficiency of supernovae feedback. Adopting a relatively strong feedback (see also Guo et al. 2011), the agreement with the observational data becomes satisfactory at the faint end. It is interesting to note, however, that matching the faint end of the luminosity function comes at the expenses of exacerbating the excess of luminous bright objects. This is due to the fact that the material reheated and/or ejected by low-mass galaxies in this model (but this is generally true for most of the models that can be found in the literature) ends up in the hot gas that is associated with the corresponding central galaxies. At later times, this material cools efficiently onto the corresponding central galaxies increasing their luminosities and star formation rates, at odds with observational data.

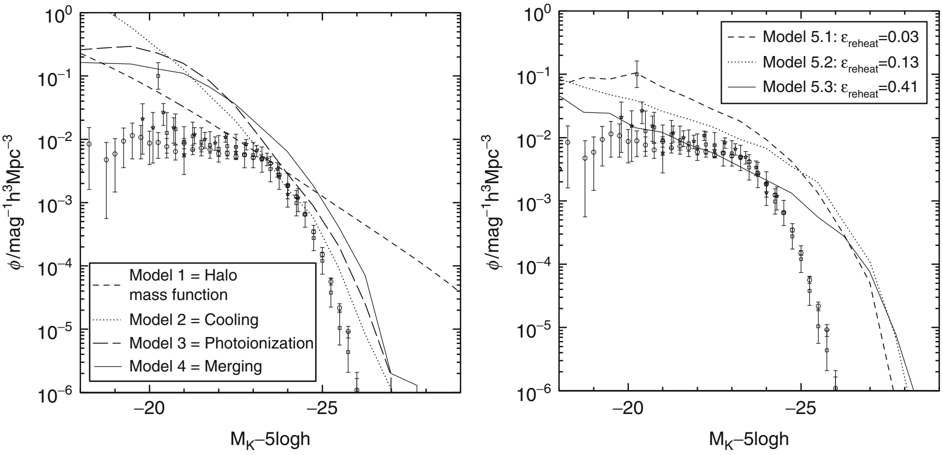


Fig. 10-20

From Benson et al. (2003). *Points* show observational determinations of the observed K-band luminosity function. *Left panel*: lines show results from different models: model 1 is obtained converting the dark matter halo mass function into a galaxy luminosity function by assuming a fixed mass-to-light ratio chosen to match the knee of the luminosity function; model 2 shows results from a model with no feedback from supernovae, no photoionization suppression, and no mergers; model 3 is similar to model 2 but includes photoionization; model 4 includes also galaxy mergers. *Right panel*: lines show predictions from a model with increasing efficiency of supernovae feedback, as indicated in the legend

Matching the bright end of the luminosity function has proved much more difficult than matching the faint end, and a reasonable success has been achieved only recently by means of a relatively strong form of "radio-mode" AGN feedback (see ➲ Sect. 3.4.3). Different prescriptions of AGN feedback have been proposed in recently published models (e.g., Croton et al. 2006; Bower et al. 2006; Monaco et al. 2007), and still much work remains to be done in order to understand if and how the energy injected by intermittent radio activity at the cluster center is able to efficiently suppress the cooling flows. In addition, recent work has pointed out that most models assume a strong dependence of radio-mode feedback on the parent halo mass. As a consequence of this assumption, these models predict that essentially all massive galaxies should be associated with a bright radio source, while observational data suggest that faint and bright radio sources are found in similar environments in equal numbers (Fontanot et al. 2011). Finally, it should be noted that although AGN feedback has received much attention in recent years, the necessity of introducing a physical process to suppress the condensation of gas at the center of massive haloes and the hypothesis that this might be due to feedback from AGN was noted in earlier work (e.g., Kauffmann et al. 1999).

The main reason for the success of the "radio-mode" AGN feedback is that it does not require star formation to be effective. As a consequence, this mode of feedback permits to suppress the luminosity of massive galaxies and, at the same time, to keep their stellar populations old, in qualitative agreement with observational data (see, e.g., De Lucia et al. 2006). The models also seem to reproduce, at least qualitatively, the observed trend for more massive ellipticals to have shorter star formation time scales. A good quantitative agreement has not been shown yet, and a detailed comparison between models and observations is complicated by large uncertainties associated to star formation histories extracted from observed galaxy spectra (see, e.g., Fontanot et al. 2009a).

In these models, ellipticals and bulges form mainly through mergers (for a detailed analysis of the contribution of different channels, see De Lucia et al. 2011). Naively, one expects very large numbers of mergers in the hierarchical scenario, where more massive systems form through the mergers of smaller units, and larger systems are expected to be made up by a larger number of progenitors. It is therefore interesting to ask how the number of progenitors varies as a function of galaxy mass. The left panel of ➲ *Fig. 10-21* shows the "effective number of stellar progenitors" of elliptical galaxies of different mass. This quantity represents a mass-weighted counting of the stellar systems that make up the final galaxy, and therefore provides a good proxy for the number of significant mergers required to assemble a galaxy of given mass. The figure shows results from a model where only mergers contribute to the formation of bulges (empty circles) and those from a model where bulges can also form through disk instability (filled symbols). The vertical dashed line indicates the threshold above which the morphology classification can be considered robust (the limit is set by the resolution of the parent N-body simulation). As expected, more massive galaxies are made up of more pieces. The number of effective progenitors is, however, less than two up to stellar masses of $\simeq 10^{11}\,M_{\odot}$, indicating that the formation of these systems typically involves only a small number of major mergers. Only more massive galaxies are built through a larger number of mergers, reaching up to $\simeq 5$ for the most massive systems. The right panel of ➲ *Fig. 10-21* shows the distribution of "formation" (top panel) and "assembly" redshifts (bottom panel) of model ellipticals. The formation redshift is defined here as the redshift when 50% of the stars that end up in ellipticals today are already formed, while the assembly redshift is defined as the redshift when 50% of the stars that end up in ellipticals today are already assembled in a single object. The right panel of ➲ *Fig. 10-21* shows that more massive galaxies are "older," albeit with a large scatter, but

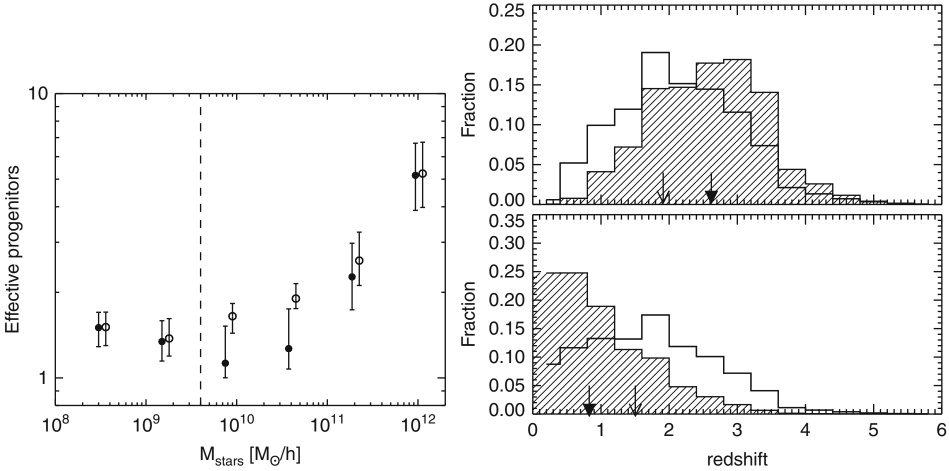


Fig. 10-21

From De Lucia et al. (2006). *Right panel*: distribution of formation (*top panel*) and assembly redshifts (*bottom panel*). The *shaded histogram* is for elliptical galaxies with stellar mass larger than 10^{11} $M_\odot$, while the open histogram is for all ellipticals with mass larger than 4 × 10^9 $M_\odot$. *Arrows* indicate the medians of the distributions, with the *thick arrows* referring to the *shaded histograms*. *Left panel*: effective number of progenitors as a function of galaxy stellar mass for model elliptical galaxies. *Symbols* show the median of the distribution, while error bars indicate the *upper* and *lower* quartiles. *Filled* and *empty symbols* refer to a model with and without a disk instability channel for the formation of the bulge

assemble "later" than their lower-mass counterparts. The assembly history of ellipticals hence parallels the hierarchical growth of dark matter haloes, in contrast to the formation history of the stars themselves. Data shown in the right panel of *Fig. 10-21* imply that a significant fraction of present elliptical galaxies have assembled relatively recently, through purely stellar mergers. This finding appears to be supported by recent observational results (e.g., van Dokkum 2005).

Models predict an increase in stellar mass by a factor 2–4 since $z \sim 1$, depending on stellar mass (De Lucia et al. 2006; De Lucia and Blaizot 2007). This creates a certain tension with the observation that the massive end of the galaxy mass function does not appear to evolve significantly over the same redshift interval. A large part of this tension is removed when taking into account observational errors and uncertainties on galaxy mass estimates (see Fontanot et al. 2009a, and references therein). For the mass assembly of the BCGs, the situation is worse: while observations seem consistent with no mass growth since $z \sim 1$, models predict an increase in mass by a factor about 4 (De Lucia and Blaizot 2007; Whiley et al. 2008). One major caveat in this comparison, however, is given by the fact that observational studies usually adopt fixed metric aperture magnitudes (which account for about 25–50% of the total light contained in the BCG and intracluster light), while models compute total magnitudes. Semianalytic models do not provide information regarding the spatial distribution of the BCG light, so aperture magnitudes cannot be calculated. In addition, most of the available models do not take into account the stripping of stars from other cluster galaxies due to tidal and harassment effects (Monaco et al. 2006; Conroy et al. 2007).

Most of the models currently available exhibit a remarkable degree of agreement with a large number of observations for the galaxy population in the local universe (e.g., the observed relations between stellar mass, gas mass, and metallicity; the observed luminosity, color, and morphology distribution; the observed two-point correlation functions). When analyzed in detail, however, some of these comparisons show important and systematic (i.e., common to most of the semianalytic models discussed in the literature) disagreements. A few of the problems on which the community is focusing in current years are discussed in the following.

Although models are not usually tuned to match observations of galaxy clustering, they generally reproduce the observed dependence of clustering on magnitude and color. The agreement appears particularly good for the dependence on luminosity, while the amplitude difference on color appears greater in the models than observed (Springel et al. 2005b). This problem might be (at least in part) related to the excess of small red satellite galaxies which plagues all models discussed in the recent literature (e.g., see Fig. 11 in Croton et al. 2006 and discussion in Fontanot et al. 2009a). At low redshift, this excess is largely due to satellite galaxies that were formed and accreted early on and that are dominated by old stellar populations. As explained in ➜ Sect. 3.7.4, semianalytic models assume that when a galaxy is accreted onto a larger structure, the gas supply can no longer be replenished by cooling that is suppressed by an instantaneous and complete stripping of the hot-gas reservoir. Since this process is usually combined with a relatively efficient supernovae feedback, galaxies that are accreted onto a larger system consume their gas very rapidly, moving onto the red sequence on quite short time scales (Weinmann et al. 2006; Wang et al. 2007). This contributes to produce an excess of faint and red satellites and a transition region (sometimes referred to as "green valley") which does not appear to be as well populated as observed. Much effort has been recently devoted to this problem, and many models have implemented a noninstantaneous stripping of the hot halo around satellites (e.g., Font et al. 2008; Weinmann et al. 2010; Guo et al. 2011). With these modifications, a larger fraction of the model satellites have bluer colors, resulting in a color distribution that is in better (but not perfect) agreement with the observational data. These models, however, still appear to overestimate the number of low- and intermediate-mass galaxies at higher redshift and the clustering signal on small scales (see, e.g., Figs. 20 and 23 in Guo et al. 2011).

The completion of new high-redshift surveys has recently pushed comparisons between model results and observational data to higher redshift (Stringer et al. 2009; de la Torre et al. 2011). This currently still rather unexplored regime for models of galaxy formation is very interesting because it is at high redshift that predictions from different models differ more dramatically.

To close this section, it is worth reminding that a long-standing problem for hierarchical models has been to match the zero point of the Tully-Fisher relation (the observed correlation between the rotation speed and the luminosity of spiral disks, Tully and Fisher 1977) while reproducing, at the same time, the observed luminosity function. As discussed in Baugh (2006), no model with a realistic calculation of galaxy size has been able to match the zero point of the Tully-Fisher relation using the circular velocity of the disk measured at the half-mass radius. It remains unclear if this difficulty is related to some approximation in the size calculation, or if it is related to more fundamental shortcomings of the cold dark matter model.

5 Concluding Remarks

This chapter has assumed that the cosmogony of our universe is well described by the ΛCDM paradigm. As a matter of fact, the CDM paradigm is not without its problems. The most discussed ones are related to the central mass distribution of low-surface brightness galaxies, to the existing substructure in galaxy-size haloes, and to the angular momentum of the galaxy disks (e.g., Tasitsiomi 2003; Benson 2005, and references therein). If there is a CDM "crisis" then, it is on the galactic and subgalactic scale, where the influence of the baryonic component is expected to play an important role. Indeed, many (all?) of the problems listed above might have an astrophysical solution, so rather than problems of the CDM scenario they might be problems with the way we model the evolution of the baryons in the cosmological context. It is clear then that in order to really test CDM, we need to improve our galaxy formation models so as to make more accurate predictions on small scales.

As discussed in ➋ Sect. 3 of this chapter, galaxy formation is a very difficult physical problem as one should account for a variety of phenomena that act on different scales and at different times and that interact in many possible ways. In addition, both theoretically and observationally, we have a very limited understanding of most of the physical processes that should be taken into account. Given the complexity of the problem, it is clear that we are not yet (and perhaps we will never be) in the position of being able to model galaxy formation "from first principles." We can, however, use a number of different techniques that can help us improving our understanding of the physical processes at play.

The theoretical tools that can be employed to study galaxy formation and evolution are many and complementary. None of them will ever provide "the model" that reproduces the observed universe. Indeed, it would be perhaps naive to believe that it is possible to summarize all the complexity that we observe in a set of analytic or semianalytic or seminumerical equations. The way to proceed then is to take advantage of the complementarity between different approaches and use the observational data to falsify the hypotheses that have been made. In going into this loop, one has to remember that the models generally include a number of free parameters. However, more than to the exact value of the parameters, attention should be given to the "parameterizations," i.e., specific assumptions on the physical processes considered. It is clear that the larger the number of processes considered is, the larger will the number of parameters/parametrizations be. There will be some degeneracy that can, however, be limited by considering a larger set of observational constraints.

These are exciting times to study galaxy formation. More and better data are becoming available. Theoretical models that try to reproduce the ever more detailed observational picture of the universe, will also require ever more complex modeling. Only by keeping the close link between theoretical predictions and observational data discussed, will it be possible to shed light on the physical processes governing galaxy formation and evolution.

Acknowledgments

The author acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n. 202781.

This work is dedicated to Luigi Cuomo.

Cross-References

- Active Galactic Nuclei
- Cosmology
- Galaxy Interactions
- Large Scale Structure of the Universe
- Star Formation
- The Influence of Environment on Galaxy Evolution

References

Abel, T., Bryan, G. L., & Norman, M. L. 2002, Science, 295, 93

Agertz, O., Moore, B., Stadel, J., Potter, D., Miniati, F., Read, J., Mayer, L., Gawryszczak, A., Kravtsov, A., Nordlund, Å., Pearce, F., Quilis, V., Rudd, D., Springel, V., Stone, J., Tasker, E., Teyssier, R., Wadsley, J., & Walder, R. 2007, MNRAS, 380, 963

Arrigoni, M., Trager, S. C., Somerville, R. S., & Gibson, B. K. 2010, MNRAS, 402, 173

Bastian, N., Covey, K. R., & Meyer, M. R. 2010, ARA&A, 48, 339

Baugh, C. M. 1996, MNRAS, 280, 267

Baugh, C. M. 2006, Rep. Prog. Phys., 69, 3101

Begelman, M. C. 2004, Coevolution of Black Holes and Galaxies (Cambridge University Press), 374

Benson, A. J. 2005, Small Scales, Big Issues for Cold Dark Matter (London: Imperial College Press), 59–75

Benson, A. J., & Bower, R. 2011, MNRAS, 410, 2653

Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M., & Lacey, C. G. 2000, MNRAS, 311, 793

Benson, A. J., Pearce, F. R., Frenk, C. S., Baugh, C. M., & Jenkins, A. 2001, MNRAS, 320, 261

Benson, A. J., Lacey, C. G., Baugh, C. M., Cole, S., & Frenk, C. S. 2002, MNRAS, 333, 156

Benson, A. J., Bower, R. G., Frenk, C. S., Lacey, C. G., Baugh, C. M., & Cole, S. 2003, ApJ, 599, 38

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Bertschinger, E. 1985, ApJS, 58, 39

Best, P. N., von der Linden, A., Kauffmann, G., Heckman, T. M., & Kaiser, C. R. 2007, MNRAS, 379, 894

Bett, P., Eke, V., Frenk, C. S., Jenkins, A., Helly, J., & Navarro, J. 2007, MNRAS, 376, 215

Bigiel, F., Leroy, A., Walter, F., Brinks, E., de Blok, W. J. G., Madore, B., & Thornley, M. D. 2008, AJ, 136, 2846

Bigiel, F., Leroy, A., & Walter, F. 2010, Computational Star Formation, Proceedings of the International Astronomical Union, IAU Symposium, 270, 327–334

Binney, J. 1977, ApJ, 215, 483

Binney, J., & Merrifield, M. 1998, Galactic Astronomy (Princeton: Princeton University Press)

Birnboim, Y., & Dekel, A. 2003, MNRAS, 345, 349

Blumenthal, G. R., Faber, S. M., Primack, J. R., & Rees, M. J. 1984, Nature, 311, 517

Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., & Lacey, C. G. 2006, MNRAS, 370, 645

Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2008, MNRAS, 383, 93

Brüggen, M., & De Lucia, G. 2008, MNRAS, 383, 1336

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582

Chabrier, G. 2003, PASP, 115, 763

Chandrasekhar, S. 1943, ApJ, 97, 255

Ciardi, B., & Ferrara, A. 2005, Space Sci. Rev., 116, 625

Clark, P. C., Glover, S. C. O., Klessen, R. S., & Bromm, V. 2011, ApJ, 727, 110

Cole, S. 1991, ApJ, 367, 45

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, ApJ, 647, 201

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2007, ApJ, 668, 826

Conroy, C., Gunn, J. E., & White, M. 2009, ApJ, 699, 486

Cowie, L. L., & Songaila, A. 1977, Nature, 266, 501

Cox, T. J., Jonsson, P., Somerville, R. S., Primack, J. R., & Dekel, A. 2008, MNRAS, 384, 386

Crain, R. A., Theuns, T., Dalla Vecchia, C., Eke, V. R., Frenk, C. S., Jenkins, A., Kay, S. T., Peacock, J. A., Pearce, F. R., Schaye, J., Springel, V., Thomas, P. A., White, S. D. M., & Wiersma, R. P. C. 2009, MNRAS, 399, 1773

Croton, D. J., Springel, V., White, S. D. M., De Lucia, G., Frenk, C. S., Gao, L., Jenkins, A., Kauffmann, G., Navarro, J. F., & Yoshida, N. 2006, MNRAS, 365, 11

Croton, D. J., Gao, L., & White, S. D. M. 2007, MNRAS, 374, 1303

Davé, R., Oppenheimer, B. D., & Finlator, K. 2011, MNRAS, 415, 11

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

de la Torre, S., Meneux, B., De Lucia, G., Blaizot, J., Le Fèvre, O., Garilli, B., Cucciati, O., Mellier, Y., et al. 2011, A&A, 525, A125+

De Lucia, G., & Blaizot, J. 2007, MNRAS, 375, 2

De Lucia, G., Kauffmann, G., Springel, V., White, S. D. M., Lanzoni, B., Stoehr, F., Tormen, G., & Yoshida, N. 2004a, MNRAS, 348, 333

De Lucia, G., Kauffmann, G., & White, S. D. M. 2004b, MNRAS, 349, 1101

De Lucia, G., Springel, V., White, S. D. M., Croton, D., & Kauffmann, G. 2006, MNRAS, 366, 499

De Lucia, G., Boylan-Kolchin, M., Benson, A. J., Fontanot, F., & Monaco, P. 2010, MNRAS, 406, 1533

De Lucia, G., Fontanot, F., Wilman, D., & Monaco, P. 2011, MNRAS, 517–+

Doroshkevich, A. G., Zel'Dovich, Y. B., & Novikov, I. D. 1967, Sov. Astron., 11, 233

Dressler, A. 1980, ApJ, 236, 351

Drory, N., Salvato, M., Gabasch, A., Bender, R., Hopp, U., Feulner, G., & Pannella, M. 2005, ApJ, 619, L131

Efstathiou, G. 1992, MNRAS, 256, 43P

Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, ApJ, 136, 748

Einasto, J. 1965, Trudy Inst. Astrofiz. Alma-Ata, 51, 87

Einasto, J., Saar, E., Kaasik, A., & Chernin, A. D. 1974, Nature, 252, 111

Evrard, A. E. 1990, ApJ, 363, 349

Fabian, A. C., Sanders, J. S., Ettori, S., Taylor, G. B., Allen, S. W., Crawford, C. S., Iwasawa, K., Johnstone, R. M., & Ogle, P. M. 2000, MNRAS, 318, L65

Farouki, R., & Shapiro, S. L. 1981, ApJ, 243, 32

Font, A. S., Bower, R. G., McCarthy, I. G., Benson, A. J., Frenk, C. S., Helly, J. C., Lacey, C. G., Baugh, C. M., & Cole, S. 2008, MNRAS, 389, 1619

Fontanot, F., De Lucia, G., Monaco, P., Somerville, R. S., & Santini, P. 2009a, MNRAS, 397, 1776

Fontanot, F., Somerville, R. S., Silva, L., Monaco, P., & Skibba, R. 2009b, MNRAS, 392, 553

Fontanot, F., Pasquali, A., De Lucia, G., van den Bosch, F. C., Somerville, R. S., & Kang, X. 2011, MNRAS, 413, 957

Forcada-Miro, M. I., & White, S. D. M. 1997, ArXiv Astrophysics e-prints

Gao, L., White, S. D. M., Jenkins, A., Stoehr, F., & Springel, V. 2004, MNRAS, 355, 819

Gao, L., Springel, V., & White, S. D. M. 2005, MNRAS, 363, L66

Gao, L., Frenk, C. S., Boylan-Kolchin, M., Jenkins, A., Springel, V., & White, S. D. M. 2011, MNRAS, 410, 2309

Gnedin, N. Y. 2000, ApJ, 542, 535

Greif, T., Springel, V., White, S., Glover, S., Clark, P., Smith, R., Klessen, R., & Bromm, V. 2011, ApJ, 737, 75

Gunn, J. E., & Gott, J. R., III 1972, ApJ, 176, 1

Guo, Q., White, S., Boylan-Kolchin, M., De Lucia, G., Kauffmann, G., Lemson, G., Li, C., Springel, V., & Weinmann, S. 2011, MNRAS, 413, 101

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Hayashi, E., & White, S. D. M. 2008, MNRAS, 388, 2

Hayashi, E., Navarro, J. F., & Springel, V. 2007, MNRAS, 377, 50

Heckman, T. M. 2002, in ASP Conf. Ser. 254, Extragalactic Gas at Low Redshift, ed. J. S. Mulchaey & J. T. Stocke (San Francisco: ASP). 292–+, Galactic Superwinds Circa 2001

Hubble, E. P. 1936, Realm of the Nebulae (New Haven: Yale University Press)

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., & Yoshida, N. 2001, MNRAS, 321, 372

Jing, Y. P., & Suto, Y. 2002, ApJ, 574, 538

Jonsson, P. 2006, MNRAS, 372, 2

Katz, N., Hernquist, L., & Weinberg, D. H. 1992, ApJ, 399, L109

Kauffmann, G., White, S. D. M., & Guiderdoni, B. 1993, MNRAS, 264, 201

Kauffmann, G., Colberg, J. M., Diaferio, A., & White, S. D. M. 1999, MNRAS, 303, 188

Kennicutt, R. C., Jr. 1998, ApJ, 498, 541

Kereš, D., Katz, N., Weinberg, D. H., & Davé, R. 2005, MNRAS, 363, 2

Klypin, A., Gottlöber, S., Kravtsov, A. V., & Khokhlov, A. M. 1999, ApJ, 516, 530

Knebe, A., Knollmann, S. R., Muldrew, S. I., Pearce, F. R., Aragon-Calvo, M. A., Ascasibar, Y., Behroozi, P. S., Ceverino, D., et al. 2011, MNRAS, 415, 2293

Komatsu, E., Smith, K. M., Dunkley, J., Bennett, C. L., Gold, B., Hinshaw, G., Jarosik, N., Larson, D., et al. 2011, ApJS, 192, 18

Lacey, C., & Cole, S. 1993, MNRAS, 262, 627

Larson, R. B. 1975, MNRAS, 173, 671

Larson, R. B. 1976, MNRAS, 176, 31

Larson, R. B., Tinsley, B. M., & Caldwell, C. N. 1980, ApJ, 237, 692

Lemson, G., & Kauffmann, G. 1999, MNRAS, 302, 111

Li, C., & White, S. D. M. 2009, MNRAS, 398, 2177

Mac Low, M.-M., & Ferrara, A. 1999, ApJ, 513, 142

Macciò, A. V., Dutton, A. A., van den Bosch, F. C., Moore, B., Potter, D., & Stadel, J. 2007, MNRAS, 378, 55

Maraston, C. 2011, Why Galaxies Care about AGB Stars II: Shining Examples and Common Inhabitants. Astronomical Society of the Pacific, 391

Maraston, C., Strömbäck, G., Thomas, D., Wake, D. A., & Nichol, R. C. 2009, MNRAS, 394, L107

Martin, C. L. 2005, ApJ, 621, 227

Mastropietro, C., Moore, B., Mayer, L., Debattista, V. P., Piffaretti, R., & Stadel, J. 2005, MNRAS, 364, 607

Mayer, L., Governato, F., & Kaufmann, T. 2008, Adv. Sci. Lett., 1, 7

McCarthy, I. G., Frenk, C. S., Font, A. S., Lacey, C. G., Bower, R. G., Mitchell, N. L., Balogh, M. L., & Theuns, T. 2008, MNRAS, 383, 593

McKee, C. F., & Ostriker, E. C. 2007, ARA&A, 45, 565

Mihos, J. C. 2004, Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution (Cambridge/New York: Cambridge University Press), 277

Mo, H., van den Bosch, F. C., & White, S. 2010, Galaxy Formation and Evolution (Cambridge/New York: Cambridge University Press)

Monaco, P., Murante, G., Borgani, S., & Fontanot, F. 2006, ApJ, 652, L89

Monaco, P., Fontanot, F., & Taffoni, G. 2007, MNRAS, 375, 1189

Moore, B., Lake, G., & Katz, N. 1998, ApJ, 495, 139

Murray, N., Quataert, E., & Thompson, T. A. 2005, ApJ, 618, 569

Nagashima, M., Lacey, C. G., Okamoto, T., Baugh, C. M., Frenk, C. S., & Cole, S. 2005, MNRAS, 363, L31

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493

Navarro, J. F., Hayashi, E., Power, C., Jenkins, A. R., Frenk, C. S., White, S. D. M., Springel, V., Stadel, J., & Quinn, T. R. 2004, MNRAS, 349, 1039

Neto, A. F., Gao, L., Bett, P., Cole, S., Navarro, J. F., Frenk, C. S., White, S. D. M., Springel, V., & Jenkins, A. 2007, MNRAS, 381, 1450

Neyman, J., & Scott, E. L. 1952, ApJ, 116, 144

Nulsen, P. E. J. 1982, MNRAS, 198, 1007

Ocvirk, P., Pichon, C., & Teyssier, R. 2008, MNRAS, 390, 1326

Okamoto, T., Gao, L., & Theuns, T. 2008, MNRAS, 390, 920

Ostriker, J. P., & Tremaine, S. D. 1975, ApJ, 202, L113

Ostriker, J. P., Peebles, P. J. E., & Yahil, A. 1974, ApJ, 193, L1

Padmanabhan, T. 1993, Structure Formation in the Universe (Cambridge/New York: Cambridge University Press)

Parkinson, H., Cole, S., & Helly, J. 2008, MNRAS, 383, 557

Peacock, J. A. 1999, Cosmological Physics (Cambridge/New York: Cambridge University Press)

Peebles, P. J. E. 1982, ApJ, 263, L1

Percival, W. J., Scott, D., Peacock, J. A., & Dunlop, J. S. 2003, MNRAS, 338, L31

Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

Quilis, V., Moore, B., & Bower, R. 2000, Science, 288, 1617

Rees, M. J., & Ostriker, J. P. 1977, MNRAS, 179, 541

Richstone, D. O. 1976, ApJ, 204, 642

Roediger, E., & Brüggen, M. 2007, MNRAS, 380, 1399

Rubin, V. C., & Ford, W. K., Jr. 1970, ApJ, 159, 379

Salpeter, E. E. 1955, ApJ, 121, 161

Saro, A., Borgani, S., Tornatore, L., Dolag, K., Murante, G., Biviano, A., Calura, F., & Charlot, S. 2006, MNRAS, 373, 397

Saro, A., De Lucia, G., Borgani, S., & Dolag, K. 2010, MNRAS, 406, 729

Scannapieco, C., Tissera, P. B., White, S. D. M., & Springel, V. 2008, MNRAS, 389, 1137

Schaye, J., Dalla Vecchia, C., Booth, C. M., Wiersma, R. P. C., Theuns, T., Haas, M. R., Bertone, S., Duffy, A. R., McCarthy, I. G., & van de Voort, F. 2010, MNRAS, 402, 1536

Schmidt, M. 1959, ApJ, 129, 243

Sheth, R. K., & Tormen, G. 2004, MNRAS, 350, 1385

Sijacki, D., & Springel, V. 2006, MNRAS, 366, 397

Silva, L., Granato, G. L., Bressan, A., & Danese, L. 1998, ApJ, 509, 103

Smith, B. D., & Sigurdsson, S. 2007, ApJ, 661, L5

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726

Springel, V., Di Matteo, T., & Hernquist, L. 2005a, MNRAS, 361, 776

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J., & Pearce, F. 2005b, Nature, 435, 629

Strickland, D. K., & Stevens, I. R. 2000, MNRAS, 314, 511

Stringer, M. J., Benson, A. J., Bundy, K., Ellis, R. S., & Quetin, E. L. 2009, MNRAS, 393, 1127

Summers, F. J., Davis, M., & Evrard, A. E. 1995, ApJ, 454, 1

Sutherland, R. S., & Dopita, M. A. 1993, ApJS, 88, 253

Tabor, G., & Binney, J. 1993, MNRAS, 263, 323

Tasitsiomi, A. 2003, Int. J. Mod. Phys. D, 12, 1157

Thomas, D. 1999, MNRAS, 306, 655

Tinker, J. L., Conroy, C., Norberg, P., Patiri, S. G., Weinberg, D. H., & Warren, M. S. 2008, ApJ, 686, 53

Toomre, A., & Toomre, J. 1972, ApJ, 178, 623

Tremonti, C. A., Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., White, S. D. M., Seibert, M., Peng, E. W., Schlegel, D. J., Uomoto, A., Fukugita, M., & Brinkmann, J. 2004, ApJ, 613, 898

Tully, R. B., & Fisher, J. R. 1977, A&A, 54, 661

Tweed, D., Devriendt, J., Blaizot, J., Colombi, S., & Slyz, A. 2009, A&A, 506, 647

van de Voort, F., Schaye, J., Booth, C. M., Haas, M. R., & Dalla Vecchia, C. 2011, MNRAS, 554–+

van den Bosch, F. C., Mo, H. J., & Yang, X. 2003, MNRAS, 345, 923

van Dokkum, P. G. 2005, AJ, 130, 2647

Viola, M., Monaco, P., Borgani, S., Murante, G., & Tornatore, L. 2008, MNRAS, 383, 777

Wang, L., Li, C., Kauffmann, G., & De Lucia, G. 2006, MNRAS, 371, 537

Wang, L., Li, C., Kauffmann, G., & De Lucia, G. 2007, MNRAS, 377, 1419

Weiner, B. J., Coil, A. L., Prochaska, J. X., Newman, J. A., Cooper, M. C., Bundy, K., Conselice, C. J., Dutton, A. A., et al. 2009, ApJ, 692, 187

Weinmann, S. M., van den Bosch, F. C., Yang, X., & Mo, H. J. 2006, MNRAS, 366, 2

Weinmann, S. M., Kauffmann, G., von der Linden, A., & De Lucia, G. 2010, MNRAS, 406, 2249

Whiley, I. M., Aragón-Salamanca, A., De Lucia, G., von der Linden, A., Bamford, S. P., Best, P., Bremer, M. N., Jablonka, P., Johnson, O., Milvang-Jensen, B., Noll, S., Poggianti, B. M., Rudnick, G., Saglia, R., White, S., & Zaritsky, D. 2008, MNRAS, 387, 1253

White, S. D. M. 1976, MNRAS, 177, 717

White, S. D. M. 1994, ArXiv Astrophysics e-prints

White, S. D. M., & Frenk, C. S. 1991, ApJ, 379, 52

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341

White, S. D. M., Frenk, C. S., & Davis, M. 1983, ApJ, 274, L1

Wojtak, R., Łokas, E. L., Gottlöber, S., & Mamon, G. A. 2005, MNRAS, 361, L1

Yoshida, N., Stoehr, F., Springel, V., & White, S. D. M. 2002, MNRAS, 335, 762

Yoshida, N., Omukai, K., Hernquist, L., & Abel, T. 2006, ApJ, 652, 6

Zeldovich, I. B., Einasto, J., & Shandarin, S. F. 1982, Nature, 300, 407

Zwicky, F. 1937, ApJ, 86, 217

11 Evolution of Active Galactic Nuclei

Andrea Merloni[1] · *Sebastian Heinz*[2]
[1]Max-Planck-Institut für Extraterrestrische Physik, Garching, Germany
[2]Astronomy Department, University of Wisconsin-Madison, Madison, WI, USA

1 A Historical Perspective on AGN Research 504
1.1 Redshift Evolution in AGN Spectral Energy Distributions 508

2 Cosmography and Demography 511
2.1 From Number Counts to Luminosity Functions 512
2.1.1 The Evolution of Radio AGN 513
2.1.2 X-Ray Surveys and the Resolution of the X-Ray Background 515
2.1.3 Optical and Infrared Studies of QSOs 520
2.1.4 Bolometric Luminosity Functions 523
2.2 AGN Clustering and the Large-Scale Structure of the Universe 528

3 Cosmology I: The Growth of Supermassive Black Holes in Galaxies 530
3.1 A Global View of the Accretion History of the Universe 531
3.1.1 The Evolution of the SMBH Mass Function 532
3.2 The AGN-Galaxy Connection 535
3.2.1 Redshift Evolution of the Scaling Relations 536

4 Cosmology II: AGN Feedback 540
4.1 Evidence and Arguments for Feedback 541
4.2 Feedback in Galaxy Clusters 542
4.2.1 The *Chandra* and *XMM-Newton* View 544
4.2.2 Estimating the Kinetic Power of a Radio Source 546
4.2.3 Cluster Radio Sources as a Population 548
4.3 Feedback in Groups 549
4.4 Radio-Mode Feedback in Galaxies 550
4.5 Quantifying the Efficiency of the Radio Mode 553
4.6 Quasar Mode Feedback 555

5 Cosmogony 559
5.1 The First Black Holes: Observational Constraints and Theoretical Ideas 559

Acknowledgments 564

References 564

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_11, © Springer Science+Business Media Dordrecht 2013

Abstract: Supermassive black holes (SMBH) lurk in the nuclei of most massive galaxies, perhaps in all of them. The tightly observed scaling relations between SMBH masses and structural properties of their host spheroids likely indicate that the processes fostering the growth of both components are physically linked, despite the many orders of magnitude difference in their physical size. This chapter discusses how we constrain the evolution of SMBH, probed by their actively growing phases, when they shine as active galactic nuclei (AGN) with luminosities often in excess of that of the entire stellar population of their host galaxies. Following loosely the chronological developments of the field, we begin by discussing early evolutionary studies, when AGN observed at various wavelengths represented beacons of light probing the most distant reaches of the universe and were used as tracers of the large-scale structure ("cosmography"). This early study turned into a more mundane enterprise of AGN "demography," once it was realized that the strong evolution (in luminosity, number density) of the AGN population hindered any attempt to derive cosmological parameters from AGN observations directly. Following a discussion of the state of the art in the study of AGN luminosity functions, we move on to discuss the "modern" view of AGN evolution, one in which a bigger emphasis is given to the physical relationships between the population of growing black holes and their environment ("cosmology"). This includes observational and theoretical efforts aimed at constraining and understanding the evolution of scaling relations, as well as the resulting limits on the evolution of the SMBH mass function. Physical models of AGN feedback and the ongoing efforts to isolate them observationally are discussed next. Finally, we touch upon the problem of when and how the first black holes formed and the role of black holes in the high-redshift universe.

Keywords: Black Hole Physics, Galaxies: active, Galaxies: clusters, Galaxies: evolution, Galaxies: nuclei, ISM: jets and outflows, quasars: general, Surveys

List of Abbreviations: *AGN*, Active galactic nucleus; *AU*, Astronomical unit; *BAL*, Broad absorption line; *BHAR*, Black hole accretion rate; *CSS*, Compact steep spectrum; *CXRB*, Cosmic X-ray background; *DM*, Dark matter; *FR I/II* Fanaroff–Riley class I/II; *GPS*, Gigahertz peak spectrum; *ICM*, Intra-cluster medium; *IGM*, Intragroup medium; *LDDE*, Luminosity-dependent density evolution; *LADE*, Luminosity and density evolution; *LF*, Luminosity function; *LLAGN*, Low-luminosity active galactic nuclei; *PLE*, Pure luminosity evolution; *PDE*, Pure Density Evolution; *QSO*, Quasi-stellar object; *SED*, Spectral energy distribution; *SFR*, Star formation rate; *SMBH*, Super-massive black hole

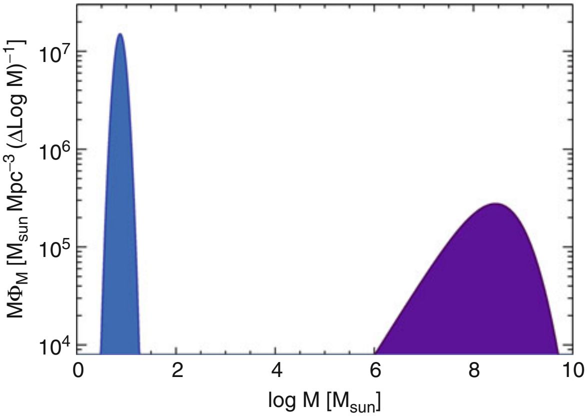
1 A Historical Perspective on AGN Research

The study of astrophysical black holes, as it has developed over the last five decades, is driven by three main rationales and goals. Because the mere existence of black holes is the most far-reaching implication of the theory of general relativity (together with the Big Bang cosmological theory), they can first of all be used to test theories of gravitation in the strong field regime. Secondly, astrophysical black holes are revealed to us through emission processes taking place in accretion flows and relativistic jets, both originating in the black hole's deep potential well, and they offer a unique opportunity of studying interesting and complex astrophysical problems, involving extreme physical conditions, relativistic magnetohydrodynamics, and radiative effects. Thirdly, black hole formation and evolution might play an important role in a broader cosmological context, affecting the formation and the evolution of the structures they live in, such as galaxies, groups, and clusters.

During the first golden age[1] of black hole astrophysics, efforts were focused on finding proof of the existence of black holes and to define their basic interactions with the environment (accretion and relativistic jet theory). Such goals only touched on the first two of the rationales listed above. The history of the development of black hole physics (both theoretical and observational) in these years has been beautifully laid out by Kip Thorne in his book *Black Holes and Time Warps: Einstein's Outrageous Legacy* (Thorne 1994), where the reader can find a more complete set of references and biographical notes, together with the historical accounts presented elsewhere in this volume (see ➲ Chap. 7).

Beginning at about the turn of the twenty-first century, black hole astrophysicists have acknowledged the relevance of their subject of study for a broader community of cosmologists and extragalactic astronomers, thanks to the multiple lines of evidence pointing toward a fundamental role played by black holes in galaxy evolution.

In fact, black holes in the local universe come in two main families according to their size, as recognized by the strongly bimodal distribution of the local black hole mass function (see ➲ *Fig. 11-1*). While the height, width, and exact mass scale of the *stellar* mass peak should be understood as a by-product of stellar (and binary) evolution and of the physical processes that make supernovae and gamma-ray bursts explode, the *supermassive* black hole peak in this distribution is the outcome of the cosmological growth of structures and of the evolution of



■ Fig. 11-1

The local black hole mass function, plotted as $M \times \phi_M$ in order to highlight the location and height of the two main peaks in the distribution. The stellar mass black hole peak has been drawn assuming a log-normal distribution with mean mass equal to 5 solar masses, width of 0.1 dex, and a normalization yielding a density of about $1.1 \times 10^7\ M_\odot$ Mpc^{-3} (Fukugita and Peebles 2004), which is about 7×10^{-5} times the critical density of the universe. The supermassive black hole peak, instead, contributes to an overall density of about $4.2 \times 10^5\ M_\odot$ Mpc^{-3} or a fraction only 2.7×10^{-6} of the critical density (see➲ Sect. 3.1 for details)

[1] This definition was introduced by the Caltech graduate Bill Press (Thorne 1994) to identify the years between the early 1960s and the early 1970s.

accretion in the nuclei of galaxies, likely modulated by the mergers these nuclear black holes will experience as a result of the hierarchical galaxy–galaxy coalescences.

This picture of the local demographics of black holes has been made possible by the discovery of tight scaling relations between the central black hole mass and various properties of their host spheroids (velocity dispersion, σ_*, stellar mass, M_*, luminosity, etc.) that characterize the structure of nearby *inactive* galaxies (Magorrian et al. 1998; Gebhardt et al. 2000; Ferrarese and Merritt 2000; Häring and Rix 2004; Gültekin et al. 2009, see ➋ *Fig. 11-2*).

These correlations are a result of the search for local QSO relics via the study of their dynamical influence on the surrounding stars and gas made possible by the launch of the *Hubble*

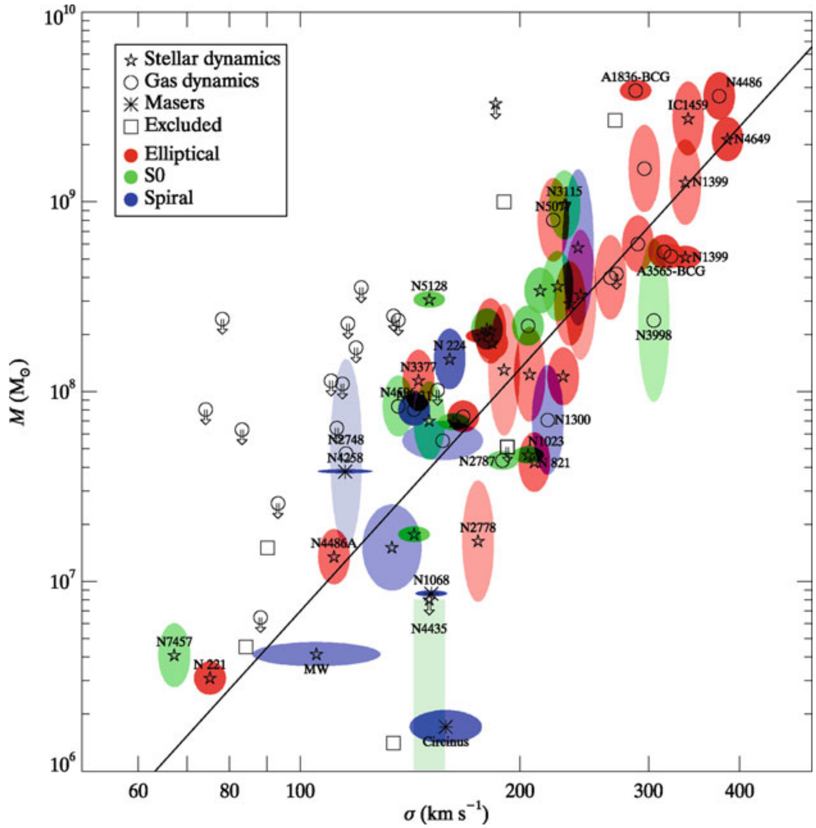


Fig. 11-2

The $M_{BH} - \sigma_*$ relation for galaxies with dynamical BH mass measurements. The *symbols* indicate the method of BH mass measurement: stellar dynamical (*pentagrams*), gas dynamical (*circles*), masers (*asterisks*). *Arrows* indicate 3σ confidence upper limits to the BH mass. The color of the error ellipse indicates the Hubble type of the host galaxy: elliptical (*red*), S0 (*green*), and spiral (*blue*). The saturation of the colors in the error ellipses or boxes is inversely proportional to the area of the ellipse or box. *Squares* are galaxies not included in the fit. This is shown as a *solid line* for the best-fit relation to the full sample: $M_{BH} = 10^{8.12}\ M_\odot(\sigma_*/200\ \ \mathrm{km\ s^{-1}})^{4.24}$ (Adopted from Gültekin et al. 2009)

Space Telescope. They have revolutionized the way we understand the physical link between the evolution of galaxies and active galactic nuclei (AGN[2]).

In addition, it is now understood that supermassive black hole (SMBH) growth is due mainly to radiatively efficient accretion over cosmological times, taking place during active phases (see ➲ Sect. 3.1 below). This, together with the understanding of a near universal presence of black holes in galactic centers, has led to the suggestion that most, if not all, galaxies went through a phase of nuclear activity in the past, during which a strong physical coupling (generally termed "feedback") might have established a long-lasting link between host and black hole properties.

Such a *renewed* interest for AGN in a cosmological context requires a good understanding of the evolutionary properties of this class of objects. The fact that AGN and quasars were a strongly evolving class of astronomical sources became evident very soon after their discovery, as we will discuss at length in the following sections. Nonetheless, the appreciation that such an evolution could not only mirror but also influence that of galaxies, groups, and clusters only became commonplace after the discovery of the above-mentioned scaling relations.

In this chapter, we will focus on the current knowledge of AGN evolution. Following loosely the chronological developments of the field, we will begin by discussing the "first generation" of AGN evolutionary studies (➲ Sect. 2), during which AGN observed at various wavelengths represented beacons of light probing the most distant reaches of the universe and were used as tracers of the structures themselves.

This short-lived epoch of AGN "cosmography" quickly gave way to a more mundane enterprise of AGN "demography," once it was realized that the strong evolution (in luminosity, number density, etc.) of the AGN population hindered any serious attempt to derive cosmological parameters from AGN observations directly. The attention then moved to the study of the *evolution* of active galactic nuclei by means of determinations of their luminosity functions. An update on the most recent works on the luminosity functions of AGN selected in different ways from different electromagnetic bands will also be given in ➲ Sect. 2, which will be closed by a brief discussion of AGN clustering as a natural complementary cosmographic tool (➲ Sect. 2.2).

We will then move to discuss the "modern" view of AGN evolution, one in which a bigger emphasis is given to the physical relationships between the population of growing black holes and their environments. We call this the "cosmology" phase of AGN studies, to highlight the close link between these subject areas that has been established in recent years. We will first discuss observational and theoretical efforts aimed at constraining and understanding the evolution of the scaling relations, as well as the resulting limits on the evolution of the SMBH mass function (➲ Sect. 3), and we will then present physical models of AGN feedback and the ongoing efforts to isolate them observationally in ➲ Sect. 4.

Finally, in the last section of this chapter (➲ Sect. 5), we will touch upon the problem of when and how the first black holes form and the role of black holes in the high-redshift universe.

Before all this, however, a small diversion is in place. To be consistent, the very notion of an evolutionary study of any particular class of objects, not only in astrophysics, requires a definition of the non-evolving *substratus* that allows us to first identify an object as a member of

[2] In this chapter, we will use both the term AGN and QSO/quasar to indicate actively growing supermassive black holes, implying no real physical distinction between the two, apart from one based on the total emitted luminosity: While AGN can be used for any objects, QSO/quasar usually identify those with bolometric luminosity $\log L_{\mathrm{bol}} > 46$ in cgs units.

the class, the evolution of which one wishes to study. To make just a simple example drawn from astronomical research, the evolution of galaxies is very much complicated by the never-ending morphological and photometric transformation of the different populations, so that a nontrivial element of any such study is the identification of progenitors and offspring along the Hubble sequence (see ❯ Chap. 1).

1.1 Redshift Evolution in AGN Spectral Energy Distributions

The overall spectral energy distribution (SED) of AGN extends over many decades in frequency and is the result of a number of different emission processes acting at different physical scales. We refer the reader to ❯ Chap. 7 in this book for a thorough discussion of these processes and of the main characteristic of AGN SED.

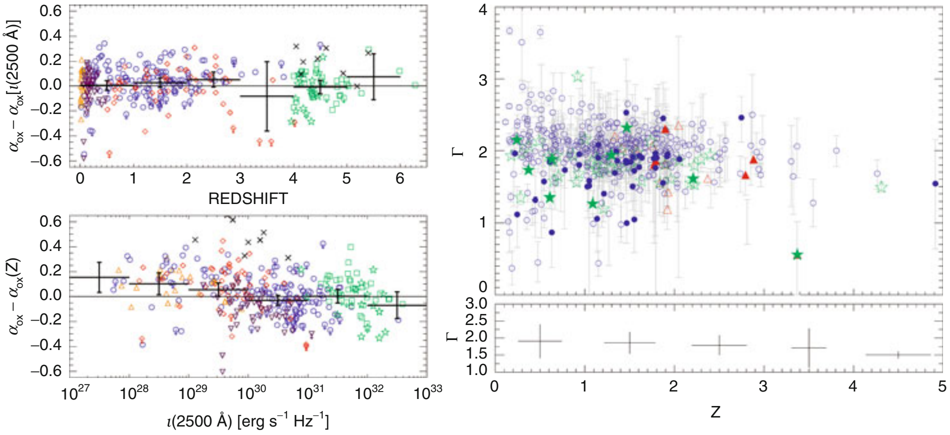
The observational appearance of an active galactic nucleus is determined not only by its intrinsic emission properties but also by the nature, amount, dynamical, and kinematic state of any intervening material along the line of sight. AGN obscuration is a crucial factor for our general understanding of the AGN phenomenon as, for example, in the traditional unification-by-orientation schemes, where different classes of AGN are explained on the basis of the line-of-sight orientation with respect to the axis of rotational symmetry of the system (see, e.g., Antonucci 1993; Urry and Padovani 1995, and references therein).

At odds with such simple schemes, evidence for a variation of the fraction of obscured AGN as a function of *luminosity* has been mounting recently (Ueda et al. 2003; Steffen et al. 2003; Simpson 2005; Hasinger 2008). The fraction of absorbed AGN, defined in different and often independent ways, appears to be lower at higher nuclear luminosities. This might be considered a signature of AGN feedback (in the “quasar” mode, see ❯ Sect. 4.6 below), in that powerful sources are able to clean up their immediate gaseous environment, responsible for the nuclear obscuration, more efficiently.

After accounting for such a clear luminosity dependence, it is currently unclear whether the overall incidence of obscuration and extinction in the nuclear regions of a galaxy evolve with redshift. This would be expected if, for example, nuclear obscuration were causally linked to the overall amount of gas within galaxies, a quantity that increases obviously with redshift.

What we are interested in here, however, is any possible evidence of redshift evolution (or lack thereof) of the *intrinsic* AGN spectral properties, that is, those characterizing the emission processes associated with the major mode of radiative energy release.

X-ray emission is ubiquitous in AGN and is a very effective way for selecting accreting black holes due both to the minimal contamination of star-forming processes and due to the decreasing importance of obscuration at increasing X-ray energies. Unfortunately, the exact mechanism responsible for AGN X-ray emission and its physical location are not fully understood yet (see ❯ Chap. 7 in this book). Still, as a very general diagnostic, the “X-ray loudness,” usually characterized by the $\alpha_{\rm ox}$ parameter, that is, the slope of the spectrum between 2,500 Å = 5 eV and 2 keV: $\alpha_{\rm ox} = 0.3838 \log(F_{2\,\rm keV}/F_{2{,}500})$ can be used to characterize the fraction of bolometric light carried away by high-energy X-ray photons. Recent studies of large samples of both X-ray and optical selected AGN have clearly demonstrated that $\alpha_{\rm ox}$ is itself a function of UV luminosity (see, e.g., Steffen et al. 2006; Young et al. 2009). However, no redshift evolution can be discerned in the data, as shown in the left panel of ❯ *Fig. 11-3*.



■ Fig. 11-3

***Left*: α_{ox} residuals as a function of redshift (*top panel*) and luminosity density at 2,500 Å (*bottom panel*). The overlaid error bars denote the mean and the 3σ standard deviation of the mean of the residuals. Limits are denoted with *arrows*. The systematic residuals in the *lower plot* indicate that α_{ox} cannot be dependent on redshift alone (adopted from Steffen et al. 2006); *Right*: X-ray photon index (Γ) versus redshift *z*. *Blue circles* represent radio quiet, non-BAL (broad absorption line) quasars *green stars* represent radio loud quasars; and *red triangles* represent BAL quasars. The *bottom plot* shows the weighted mean Γ values for bins of width Δ*z* = 1. No clear sign of evolution in the average X-ray spectral slope of AGN is detected over more than 90% of the age of the universe (From Young et al. 2009)**

Moreover, large collecting-area X-ray telescopes allow a more precise determination of the X-ray spectra of AGN, which are usually characterized by a power law, upon which emission lines and absorption features are superimposed. Up to the highest redshift where reliable spectral analysis of AGN can be performed, no clear sign of evolution in the X-ray spectral slope Γ has been detected (see the right panel of ❯ *Fig. 11-3*).

Similarly, while the narrow iron Kα emission line, the most prominent feature in AGN X-ray spectra, is clearly dependent on luminosity (the so-called Iwasawa–Taniguchi effect; (1993)), it shows *no* sign of evolution in its equivalent width with redshift, at least up to $z \simeq 1.2$ (❯ *Fig. 11-4*; Chaudhary et al. 2010, and references therein).

Even more surprising is the lack of evolution in the *optical* emission line properties of QSOs. The metallicities implied by the relative strength of broad emission lines do not show any significant redshift evolution: They are solar or super-solar, even in the highest redshift QSOs known (see, e.g., Hamann and Ferland 1992), in contrast with the strong evolution of the metallicity in star-forming galaxies.

A pictorial view of this surprising uniformity is shown in the left panel of ❯ *Fig. 11-5*, where the raw spectra from ~17,000 QSOs extracted from the Sloan Digital Sky Survey (SDSS) are plotted next to each other in a sequence of increasing redshift from bottom to top. The right panel of ❯ *Fig. 11-5* shows a direct comparison of stacked QSO spectra in three redshift intervals (Juarez et al. 2009), where it is clear that the flux ratios among the most prominent lines stay almost constant up to the highest redshift probed.

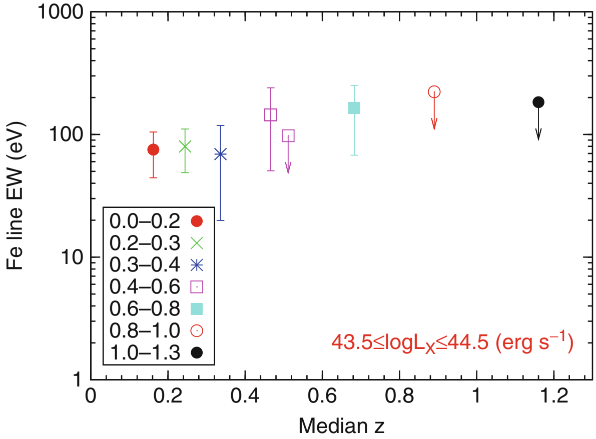


Fig. 11-4

Rest-frame equivalent width of the narrow Iron K*α* emission line observed in the average spectrum of AGN as a function of redshift. Only objects in a fixed (2–10 keV) luminosity range $10^{43.5} < L_{2-10} < 10^{44.5}$ ergs s^{-1} were considered (From Chaudhary et al. 2010)

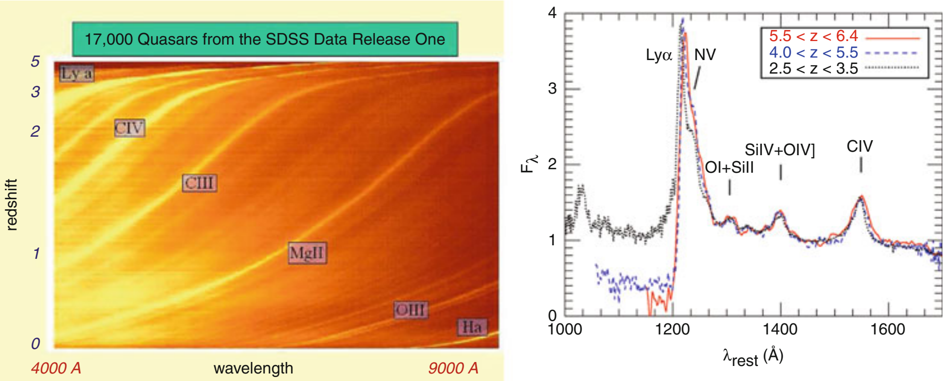


Fig. 11-5

***Left*: Spectra of 17,000 QSOs from SDSS. Notice the large degree of uniformity in the relative intensity of the main emission features (courtesy of X. Fan); *Right*: Stacked spectra of quasars in different redshift bins. Note that the relative intensity of the metal lines (and in particular the (SiIV+OIV])/CIV ratio) remains constant over the wide redshift interval 2.5 < *z* < 6.4, indicating that the metallicity in the observed quasars does not evolve with redshift (From Juarez et al. (2009))**

Summarizing, there exists a remarkably uniform set of spectral characteristics that defines active nuclei at all epochs in the history if the universe, at least if we consider objects of a fixed total (bolometric) luminosity. The simplest explanation is that the *emission* properties from AGN, that is, those which are (in most cases, at least) set by physical processes taking place within the gravitational sphere of influence of the central black hole, are essentially dictated by the gas and plasma dynamics there, where the central object's gravity dominates. We should then expect them to be relatively insensitive to the cosmological epoch, which instead greatly affects the properties of matter (density, temperature, ionization state, etc.) at the generic outer boundary, that is, right outside the SMBH gravitational sphere of influence.

2 Cosmography and Demography

Accreting supermassive black holes have long been the lighthouses of our observable universe, holding the record of the most distant object known for more than four decades. As such, they have played a key role in the early phases of cosmological investigations.

Already in 1955, the second Cambridge catalog (2C) of unresolved radio sources (the so-called radio stars) observed at 81 MHz (3.7 m) had shown both a remarkable *uniformity* in the distribution of objects in the sky and an *increase* in the cumulative number counts (see below) that allowed Ryle and Scheuer (1955) to unambiguously demonstrate not only their extragalactic origin but also that the bulk of the sources should lie at distances larger than a few tens of Mpc, that is, well beyond the edge of the optically observable universe at the time.

The dispute over the exact shape of the radio sources number count distribution that ensued soon afterward became a key part of the debate between "steady state" and "evolutionary" models of the universe, lending strong support against stationary universe models (see *Fig. 11-6*).

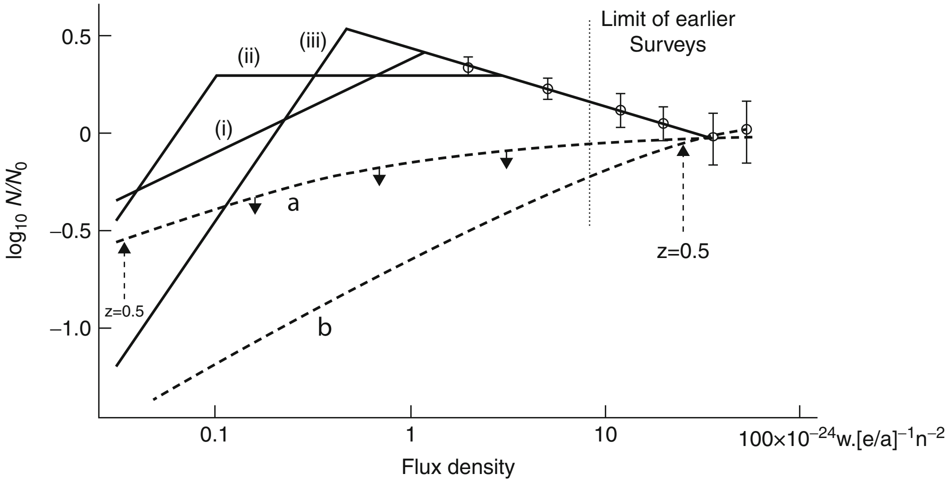


Fig. 11-6

Observed normalized radio number counts from the original 3C catalog at 178 MHz (from Ryle and Clarke 1961). The observational points (*open circles*) are extrapolated at low fluxes with three empirical models (i, ii, and iii) made such as not to violate the total low-frequency radio background available at the time. *Dashed lines* marked with (a) and (b) denote the counts predicted by the steady state cosmological model assuming two possible luminosity functions, given the observed sources. A clear discrepancy emerged between observations and non-evolving universe models. A more detailed discussion of current constraints on radio sources number counts is given in Sect. 2.1.1

Just 2 years after the discovery of quasars (Schmidt 1963) with their exceedingly large redshifts, A. Sandage wrote: *"The objects would seem to be of major importance in the solution of the cosmological problem. They can be found at great distances because of their high luminosity. Studies of the [number counts] curves using [quasars] should eventually provide a crucial test of various cosmological models"* (Sandage 1965). Similar hopes were expressed by Longair a few months later (Longair 1966).

However, as we will discuss in more detail in the following sections, quasars and radio galaxy source counts demonstrated clearly that the populations being studied *did* evolve strongly with cosmic epoch: The number of quasars per unit comoving volume was clearly larger in the past, so that the information about the geometry of the universe and the cosmological parameters is buried underneath that about the evolution of the AGN themselves.

Progress in characterizing the intrinsic evolution of the QSO population effectively quenched the hope to use black holes as ideal tracers of the structure of the universe, but opened up the study of the evolution of growing supermassive black holes, that we outline below.

2.1 From Number Counts to Luminosity Functions

By *number counts*, one typically means the surface density in the sky of a given class of sources as a function of the limiting flux of the observations. In astronomy, this is the simplest observational tool that can be used to study the evolution of a sample of objects (and to test cosmological models).

The space density of sources of different intrinsic luminosities, L, is described by the *luminosity function* (LF), $\phi(L)$, so that $dN = \phi(L)dL$ is the number of sources per unit volume with luminosity in the range L to $L + dL$. Let us consider, for simplicity, the local or nearby (Euclidean) universe uniformly filled with sources with LF $\phi(L)$. If S is the limiting flux that we can detect, sources with luminosity L can be observed out to a distance $r = (L/4\pi S)^{1/2}$. The number of sources over the solid angle Ω, observable down to the flux S are

$$N(>S) = \int \frac{\Omega}{3} r^3 \phi(L) dL = \frac{\Omega}{3(4\pi)^{3/2}} S^{-3/2} \int L^{3/2} \phi(L) dL . \qquad (11.1)$$

Thus, independent of the exact *shape* of the luminosity function entering in the determination of a normalization constant, the *slope* of the cumulative number counts of any non-evolving class of sources in a uniform, Euclidean universe should always be equal to $d \log N(>S)/d \log S = -3/2$ (if we use magnitudes, m, instead of luminosities, then d log $N(>m)/dm = 0.6$).

In general, the correct relativistic expression for number counts differs from the Euclidean one because (a) the observed flux density depends upon the spectrum of the source, as the radiation emitted at frequency ν_1 is observed at the redshifted frequency $\nu_0 = \nu_1/(1+z)$, and (b) curvature effects modify the volume element per unit redshift, making it smaller with increasing z. Overall, for typical source spectra which are not too strongly "inverted" (i.e., with flux density increasing with increasing frequency), the combination of these effects makes it more and more difficult to detect sources at progressively higher redshift and causes number counts to have slopes always shallower than the Euclidean one (see, e.g., Longair 2008, Chapter 17). As we will see below, strong evolutionary effects (i.e., luminosity functions changing rapidly with time) can counteract such a behavior.

Before proceeding, a brief introduction of common terminology widely adopted in the study of luminosity function evolution is necessary. The simplest and most general approach describes an evolving luminosity function with the aid of two functions, $f_{\rm l}(z)$ and $f_{\rm d}(z)$, that take into account the evolution of the luminosity and number density of the sources, respectively:

$$\phi(L, z) = f_{\rm d}(z)\phi(L/f_{\rm l}(z), z = 0). \tag{11.2}$$

In the *pure luminosity evolution (PLE)* case ($f_{\rm d}$ = const.), the comoving number density of sources is constant, but luminosity varies with cosmic epoch; in the *pure density evolution (PDE)* case ($f_{\rm l}$ = const.), but the comoving density of sources of any luminosity varies.

In the following sections, we will discuss the observational state of the art as far as AGN number counts and luminosity functions are concerned, in the radio, X-rays, and optical/IR bands. More comprehensive and specialized reviews have, of course, been published. In particular, we refer to the recent work by de Zotti et al. (2010) for a discussion of observations at radio wavelengths, Croom et al. (2009) for optical QSOs, and to Brandt and Hasinger (2005) for X-ray studies.

2.1.1 The Evolution of Radio AGN

➲ *Figure 11-7* shows a compilation of cumulative source number counts from a large number of surveys in different radio bands (data points from Massardi et al. 2010, see references therein). On the bottom x-axis, the total radio flux is expressed as $S_{\rm R} \equiv \nu S_\nu$ (where S_ν is the observed radio flux density at any given radio frequency ν) in cgs units, while the top axis shows the corresponding radio flux density at 1.4 GHz. Overall, the shape of the radio counts is similar in all bands, indicating the relative lack of spectral complexity of radio AGN. This is best seen when normalizing the observed counts to the Euclidean slope, as shown in the top panel.

At bright fluxes, counts rise more steeply than $S^{-3/2}$. This was already discovered by the first radio surveys at meter wavelengths (Ryle and Scheuer 1955), as we have discussed above, lending strong support for evolutionary cosmological models, as opposed to theories of a steady state universe (see ➲ *Fig. 11-6*).

At fluxes fainter than about a Jansky[3] (or ≈10^{-14} ergs s^{-1} cm^{-2} at 1 GHz), the counts increase less steeply than $S^{-3/2}$, being dominated by sources at high redshift, thus probing a substantial volume of the observable universe.

At flux densities above one mJy, the population of radio sources is largely composed by AGN. For these sources, the observed radio emission includes the classical extended jet and double lobe radio sources as well as compact radio components more directly associated with the energy generation and collimation near the central engine.

The deepest radio surveys, however, (see, e.g., Padovani et al. 2009 and references therein), probing well into the sub-mJy regime, clearly show a further steepening of the counts. The nature of this change is not completely understood yet, but in general, it is attributed to the emergence of a new class of radio sources, most likely that of star-forming galaxies and/or radio quiet AGN. Unambiguous solutions of the population constituents at those faint flux levels require not only identification of the (optical/IR) counterparts of such faint radio sources but also a

[3] A Jansky (named after Karl Jansky, who first discovered the existence of radio waves from space) is a flux measure, corresponding to 10^{-23} ergs cm^{-2} Hz^{-1}.

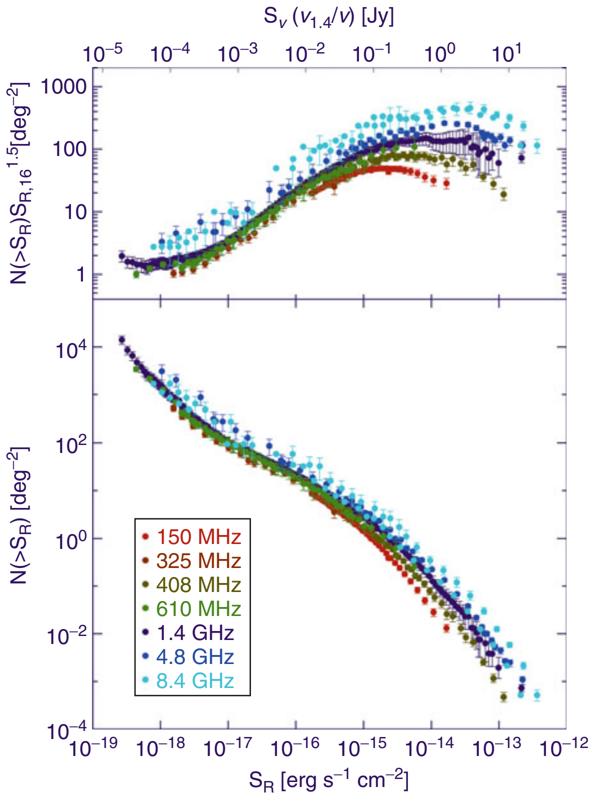


Fig. 11-7

A compilation of cumulative radio source counts (number of objects brighter than a given flux per square degree) in various bands. The observational data are taken from (Massardi et al. 2010) (see also references therein). On the *bottom horizontal axis* the total radio flux $S_R = \nu S_\nu$ in CGS units is shown, while the *top horizontal axis* shows the corresponding flux density in Jansky, where $\nu_{1.4}$ is the frequency of 1.4 GHz. The *bottom panel* shows the observed counts, while the *top panel* shows the counts after the Euclidean slope has been factored out

robust understanding of the physical mechanisms responsible for the observed emission both at radio and optical/IR wavelengths.

Thus, the complex shape of the observed number counts provides clues about the evolution of radio AGN, as well as on their physical nature, even before undertaking the daunting task of identifying substantial fractions of the observed sources, determining their distances, and translating the observed density of sources in the redshift-luminosity plane into a (evolving) luminosity function. Pioneering work from Longair (1966) already demonstrated that, in order to reproduce the narrowness of the observed "bump" in the normalized counts around 1 Jy seen in ➋ *Fig. 11-7*, only the most luminous sources could evolve strongly with redshift. This was probably the first direct hint of the intimate nature of the *differential* evolution AGN undergo over cosmological times.

Indeed, many early investigations of high-redshift radio luminosity functions (see, e.g., Danese et al. 1987) demonstrated that neither PLE nor PDE models could explain the observed evolution of radio sources, with more powerful sources (often of FRII morphology, see ➋ Chap. 7) displaying a far more dramatic rise in their number densities with increasing redshift (see also Willott et al. 2001).

Trying to assess the nature of radio AGN evolution across larger redshift ranges requires a careful evaluation of radio spectral properties of AGN. Steeper synchrotron spectra are produced in the extended lobes of radio jets, while flat spectra are usually associated with compact cores. For objects at distances such that no radio morphological information is available, the combination of observing frequency, K-corrections, intrinsic source variability, and orientation of the jet with respect to the line of sight may all contribute to severe biases in the determination of the comoving number densities of sources, especially at high redshift (Wall et al. 2005).

In a very extensive and equally influential work, Dunlop and Peacock (1990) studied the evolution of the luminosity functions of steep, and flat-spectrum sources separately. They showed that the overall redshift evolution of the two classes of sources were similar, with steep-spectrum sources outnumbering flat ones by almost a factor of 10. Uncertainties remained regarding the possibility of a high-redshift decline of radio AGN number densities. The issue is still under discussion, with the most clear evidence for such a decline observed for flat-spectrum radio QSO at $z > 3$ (Wall et al. 2005), consistent with the most recent findings of optical and X-ray surveys (see also ➋ Sect. 5 below).

Under the simplifying assumption that the overall radio AGN population can be subdivided into steep- and flat-spectrum sources, characterized by a power-law synchrotron spectrum $S_\nu \propto \nu^{-\alpha}$, with slope $\alpha_{\mathrm{flat}} = 0.1$ and $\alpha_{\mathrm{steep}} = 0.8$, respectively, a redshift-dependent luminosity function can be derived for the two populations separately, by fitting simple models to a very large and comprehensive set of data on multifrequency source counts and redshift distributions obtained by radio surveys at $\nu < 5$ GHz (Massardi et al. 2010). The comoving number densities in bins of increasing radio power (at 1.4 GHz) from the resulting best-fit luminosity function models are shown in the left panel of ➋ *Fig. 11-8*.

Radio AGN, both with steep and flat spectrum, show the distinctive feature of a differential density evolution, with the most powerful objects evolving more strongly toward higher redshift, a phenomenological trend that, in the current cosmologist jargon, is called "downsizing."

Recent radio observational campaigns of large multiwavelength sky surveys have also corroborated this view, by providing a much more detailed picture of low-luminosity radio AGN. For example, the work of Smolčić et al. (2009) on the COSMOS field showed that radio galaxies with $L_{1.4\mathrm{GHz}} <$ few $\times 10^{25}$ WHz^{-1} evolve up to $z \simeq 1$, but much more mildly than their more luminous counterparts, as shown in the right panel of ➋ *Fig. 11-8*.

2.1.2 X-Ray Surveys and the Resolution of the X-Ray Background

As already mentioned, active galactic nuclei are powerful X-ray emitters. The discovery of the intense cosmic X-ray background radiation (CXRB; Giacconi et al. 1962) in the early 1960s opened up a privileged window for the study of the energetic phenomena associated with accretion onto black holes.

Due to the relative weakness of X-ray emission from stars and stellar remnants (magnetically active stars, cataclysmic variables, and, more importantly, X-ray binaries are the main

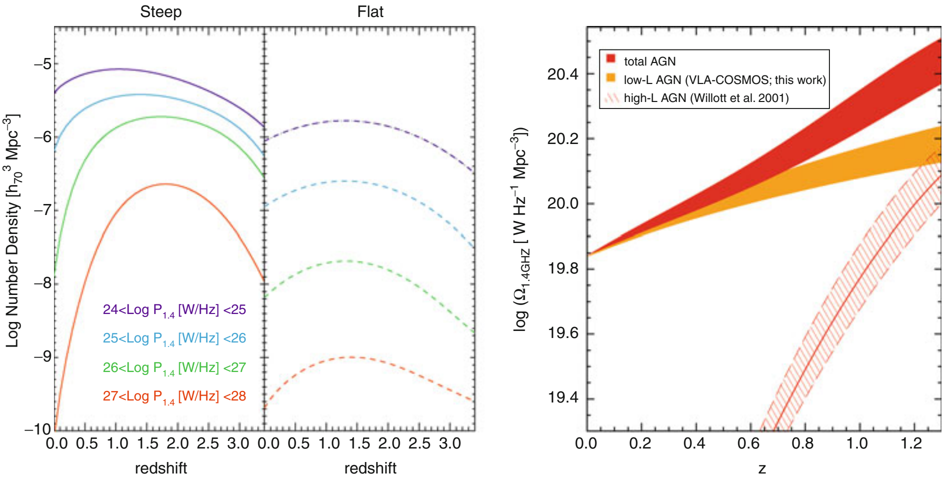


Fig. 11-8

The radio view of AGN downsizing. *Left*: Best fit number density evolution of radio sources of different power, taken from the models of Massardi et al. (2010), for steep, and flat-spectrum sources in the *left* and *right panels*, respectively. *Right*: Evolution of the comoving 20-cm integrated luminosity density for VLA-COSMOS AGN (*orange curve*) galaxies for $z < 1.3$. Also shown is the evolution of the high-luminosity radio AGN, adopted from Willott et al. (2001) (hatched region; the *thick* and *dashed lines* correspond to the mean, maximum, and minimum results, respectively). The evolution for the total AGN population, obtained by co-adding the VLA-COSMOS and high-luminosity AGN energy densities, is shown as the *red-shaded curve* (Adopted from Smolčić et al. 2009)

stellar X-ray sources), the X-ray sky is almost completely dominated by the evolving SMBH population, at least down to the faintest fluxes probed by current X-ray focusing telescopes. The goal of reaching a complete census of evolving AGN and thus of the accretion power released by SMBH in the history of the universe has therefore been intertwined with that of fully resolving the CXRB into individual sources. Accurate determinations of the CXRB intensity and spectral shape, coupled with the resolution of this radiation into individual sources, allow very sensitive tests of how the AGN luminosity and obscuration evolve with redshift.

New generations of synthesis models of the CXRB (Gilli et al. 2007; Treister et al. 2009) have quickly followed the publication of increasingly larger and deeper surveys (for the current deepest view of the X-ray sky, see *Fig. 11-9*). *Figure 11-10* shows a recent compilation of the CXRB measurements together with one incarnation of a synthesis model (Treister et al. 2009) of AGN evolution that explains those data. The hard slope of the background spectrum (well described by a power law with photon index $\Gamma_{\mathrm{CXRB}} \simeq 1.4$ at $E < 10$ keV) and the prominent peak at about 30 keV are accounted for by assuming that the majority of active galactic nuclei are in fact obscured.

These new models have progressively reduced the uncertainties in the absorbing column density distribution. When combined with the observed X-ray luminosity functions, they provide an almost complete census of the Compton-thin AGN (i.e., those obscured by columns $N_{\mathrm{H}} < \sigma_{\mathrm{T}}^{-1} \simeq 1.5 \times 10^{24}\,\mathrm{cm}^{-2}$, where σ_{T} is the Thomson cross section). This class of objects

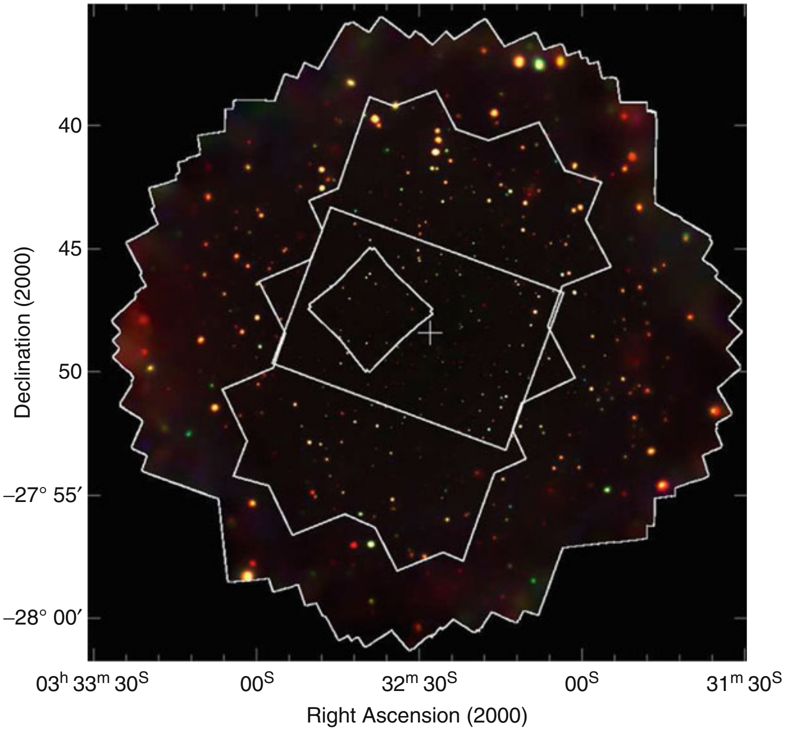


Fig. 11-9
A smoothed, false color image of the deepest X-ray exposure to date, the 4 × 10^6 s *Chandra* observation of the Chandra Deep Field South (CDFS). These observations resolve almost the entire CXRB radiation below ~5 keV into individual sources, the vast majority of which are accreting supermassive black holes (From Xue et al. 2011)

dominates the counts in the lower energy X-ray energy band, where almost the entire CXRB radiation has been resolved into individual sources (Worsley et al. 2005). It should be noted, however, that at the peak energy of the observed CXRB radiation, only a small fraction (less than 5%) of the emission has so far been resolved into individual objects.

CXRB synthesis models, like the one shown in *Fig. 11-10*, ascribe a substantial fraction of this unresolved emission to heavily obscured (Compton thick) AGN. However, because of their faintness even at hard X-ray energies, their redshift and luminosity distribution is essentially unknown, and even their absolute contribution to the overall CXRB sensitively depends on the quite uncertain normalization of the unresolved emission at hard X-ray energies. The quest for the physical characterization of this "missing" AGN population, most likely dominated by Compton-thick AGN, represents one of the last current frontiers of the study of AGN evolution at X-ray wavelengths.

Putting together the observational data from a large suite of complementary surveys, *Fig. 11-11* shows a compilation of X-ray number counts, for both soft (0.5–2 keV) and hard (2–10 keV) selected samples.

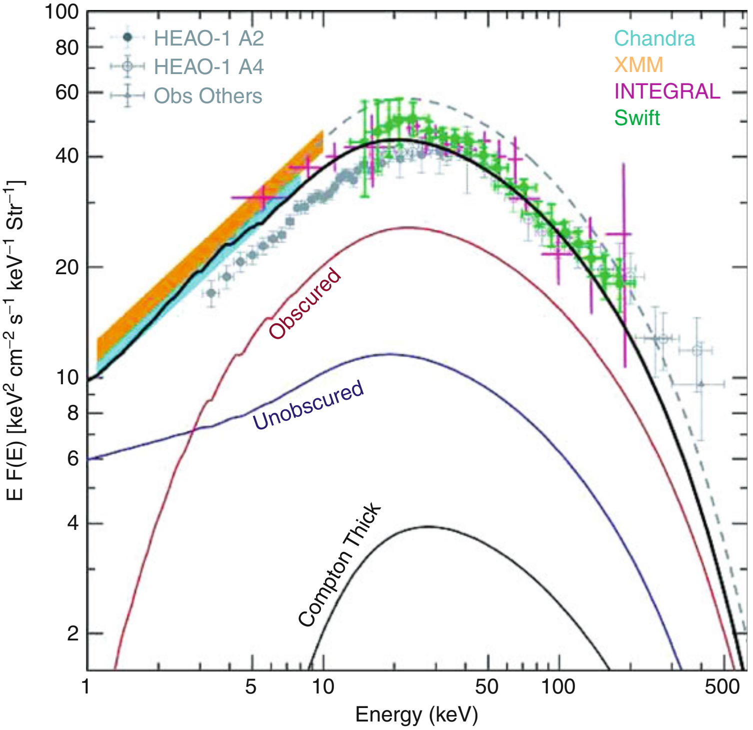


Fig. 11-10
Observed spectrum of the extragalactic CXRB from *HEAO 1*, *Chandra*, *XMM-Newton*, *INTEGRAL*, and *Swift* data. The *solid (red, blue, and black) thin lines* show the contribution to this model from unobscured and obscured Compton-thin and Compton thick AGNs, respectively. The *thick black solid* and *dashed gray lines* are two different CXRB spectral models proposed in the literature depending on the assumed normalization of the *HEAO 1*; the main difference is the number of Compton-thick AGNs, which is reduced by a factor of 4 if the *black solid line* is assumed instead of the dashed one (Adopted from Treister et al. (2009), where a complete list of references of the data points shown is presented)

Given the steep frequency dependence of photoelectric absorption cross sections, the harder the energy band probed, the less affected by obscuration the objects under study are. Current technologies provide the best compromise between telescope effective area and energy range in the 2–10-keV band. Indeed, the density of AGN detected in this band is higher than that of 0.5–2-keV selected ones, by at least a factor of 3.

As for the radio log N-log S of ➲ *Fig. 11-7*, the counts become shallower than the Euclidean slope at intermediate fluxes (about 10^{-14} ergs s^{-1} cm^{-2}), where the largest relative fraction of the CXRB is produced. Tentative evidence of a steepening at the lowest fluxes might indicate the emergence of a different, non-AGN, population of star-forming galaxies whose X-ray emission is primarily due to stars and stellar remnants.

The deepest surveys so far carried out in the soft X-ray energy range (0.5–2 keV), supplemented by the painstaking work of optical identification and redshift determination of the

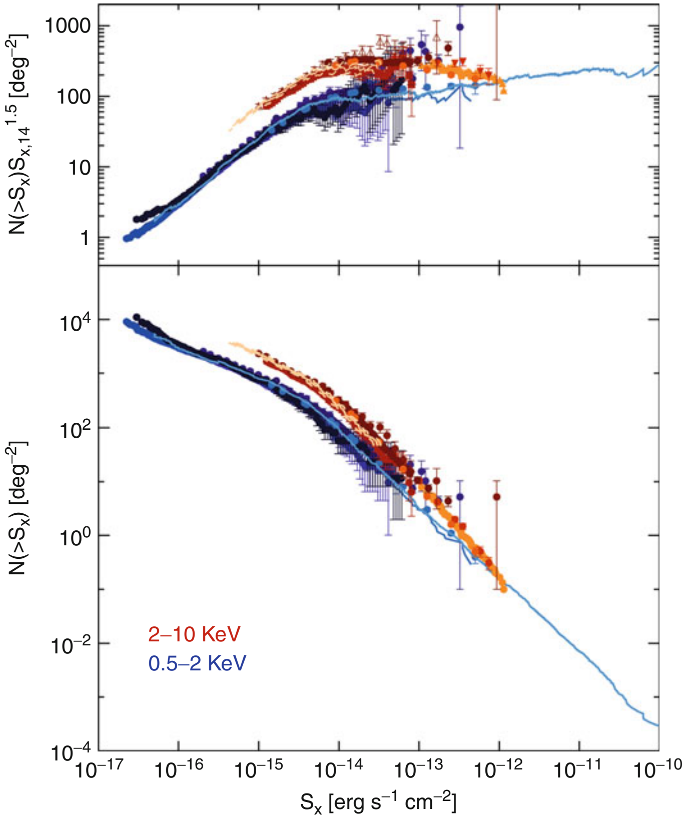


Fig. 11-11

A compilation of cumulative X-ray source counts (number of objects brighter than a given flux per square degree) in the soft (0.5–2 keV, *cyan-blue colors*) and hard (2–10 keV, *red-orange colors*) bands. The observational data are taken from Cappelluti et al. (2009, see references therein) and Mateos et al. (2008). The *bottom panel* shows the observed counts, while the *top panel* shows the counts after the Euclidean slope has been factored out

detected sources have, provided the most accurate description of the overall evolution of the AGN luminosity function. Neither PLE nor PDE provides a satisfactory description of the X-ray LF evolution, with a good fit to the data achieved with a "Luminosity dependent density evolution" (LDDE) model or variations thereof. In their influential work, Hasinger et al. (2005) unambiguously demonstrated that in the observed soft X-ray energy band, more luminous AGN peaked at higher redshift than lower luminosity ones (see ➋ *Fig. 11-12*).

A good enough sampling of the luminosity redshift plane necessary for accurate LF studies requires more extensive observational efforts in the hard X-ray band, as obscured AGN are more difficult to identify (and to obtain redshifts for) in the optical band. Nonetheless, the general "downsizing" trend illustrated by the soft-X-ray selected AGN of ➋ *Fig. 11-12* has so far been confirmed by almost all recent studies of (2–10 keV) X-ray selected AGN (see, e.g., Ueda et al. 2003; Barger et al. 2005).

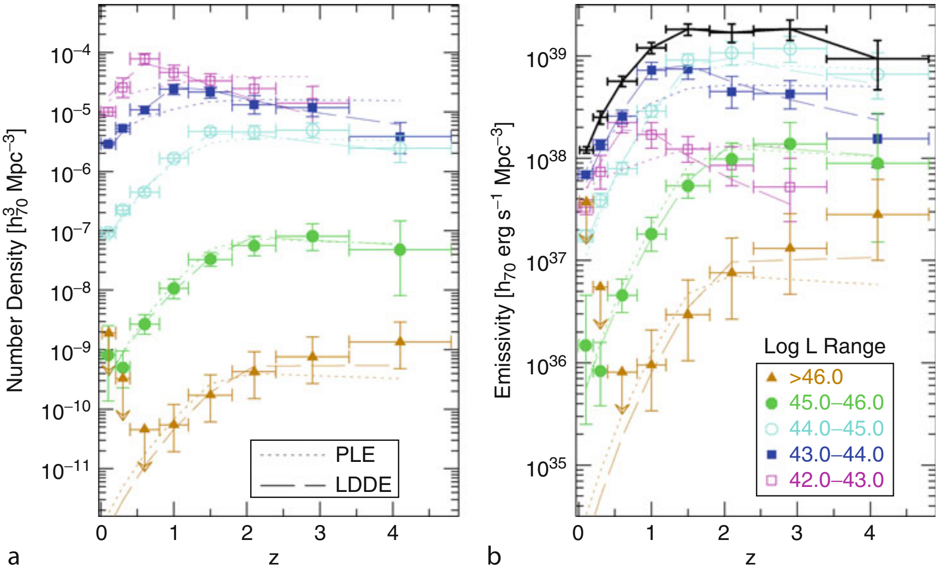


Fig. 11-12

The soft X-ray view of AGN downsizing: (a) the space density of AGNs as a function of redshift in different luminosity classes and the sum over all luminosities with log $L_x \gtrsim 42$. Densities from the PLE and LDDE models are overplotted with *solid lines*. (b) The same as (a), except that the soft X-ray emissivities are plotted instead of number densities. The *uppermost curve* (*black*) shows the sum of emissivities in all luminosity classes in the plot (From Hasinger et al. (2005))

2.1.3 Optical and Infrared Studies of QSOs

Bright AGN emit a large fraction of their bolometric luminosity in the optical/UV part of the spectrum (see Chap. 7). For Eddington ratios ($\lambda \equiv L_{\rm bol}/L_{\rm Edd}$, where $L_{\rm Edd} = 4\pi G M_{\rm BH} m_{\rm p} c/\sigma_{\rm T} \simeq 1.3 \times 10^{38} (M_{\rm BH}/M_{\odot})$ ergs s^{-1} is the Eddington luminosity) larger than a few percent, the AGN light outshines the emission from the host galaxy, resulting in point-like emission with peculiar blue colors.

Finding efficient ways to select QSO in large optical surveys, trying to minimize contamination from stars, white dwarfs, and brown dwarfs has been a primary goal of optical astronomers since the realization that QSO were extragalactic objects often lying at cosmological distances (Schmidt and Green 1983; Richards et al. 2006).

Optical surveys remain an extremely powerful tool to uncover the evolution of unobscured QSOs up to the highest redshift ($z \sim 6$). In terms of sheer numbers, the known population of SMBH is dominated by such optically selected AGN (e.g., more that 10^5 QSOs have been identified in the Sloan Digital Sky Survey), essentially due to the yet unsurpassed capability of ground-based optical telescopes to perform wide-field, deep surveys of the extragalactic sky.

Figure 11-13 shows a compilation of (g-band, ≈4, 700Å) quasar number counts from the largest recent optical surveys (2SLAQ, 2QZ, and SDSS), for objects in the redshift range $0.3 < z < 2.2$. The overall shape is similar to that of the radio AGN counts, with a steep increase at bright fluxes, followed by a turnover at around $g \simeq 19$. A comparison with *Figs. 11-7* and *11-11* reveals, however, that such large area QSO surveys reach depths corresponding to

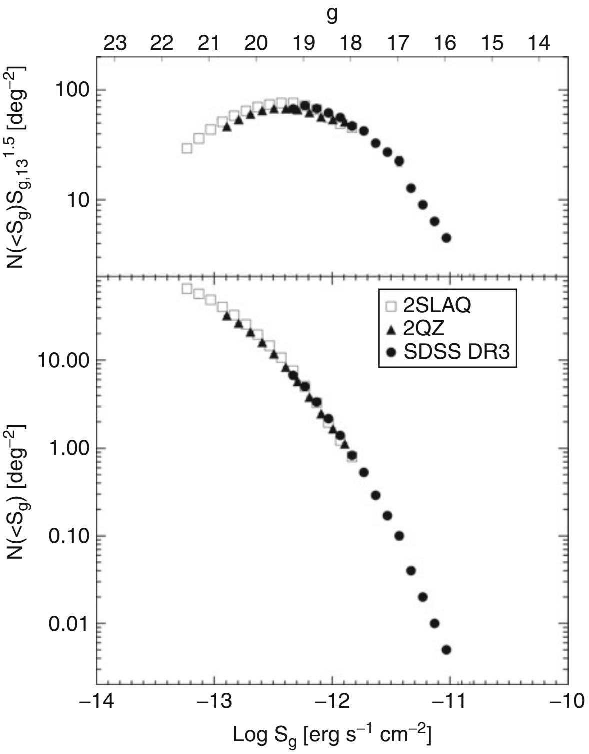


Fig. 11-13

Compilation of (g-band, ≈4, 700 Å) quasar number counts from the largest recent optical surveys at $0.3 < z < 2.3$ (2SLAQ, 2QZ and SDSS, courtesy of G. Richards and S. Croom). g-band magnitudes have been converted into CGS flux units (S_g) with $\log S_g = -0.4(g - 20) + 3.359$

a number density of sources in the sky more than one order of magnitude smaller that those probed by the deepest X-ray and radio surveys.

Thus, the dominant AGN population eludes systematic detection in optical surveys. In general terms, the difficulty of optical QSO surveys to probe deep into the AGN population is due to two major effects: The first one is the already mentioned issue of *nuclear obscuration*, dramatically affecting the UV/optical appearance of AGN; the second is *galaxy dilution* of the AGN light (and of the broad emission line signature often used to select quasars). More specifically, let us consider an AGN with B-band luminosity given by $L_{\rm AGN,B} = \lambda L_{\rm Edd} f_{\rm B}$, with bolometric correction $f_{\rm B} \approx 0.1$ (Richards et al. 2006). Assuming a bulge-to-black hole mass ratio of 0.001 and a bulge-to-total galactic stellar mass ratio of (B/T), the contrast between nuclear AGN continuum and host galaxy blue light is given by

$$\frac{L_{\rm AGN,B}}{L_{\rm host,B}} = \frac{\lambda}{0.1}\frac{(M_*/L_{\rm B})_{\rm host}}{3(M_\odot/L_\odot)}(B/T). \tag{11.3}$$

Thus, for typical mass-to-light ratios, the AGN will become increasingly diluted by the host stellar light at Eddington ratios λ smaller than a few percent.

High-spatial resolution observations of the numerically dominant population of Low-luminosity AGN (LLAGN; see the comprehensive review of Ho 2008, and references therein) have so far only been possible in the very local universe. At higher redshift, the deepest multiwavelength AGN-galaxy surveys to date are starting to probe AGN luminosities such that the contribution of the host galaxy to the overall SED cannot be neglected. This compromises the efficiency and "cleanness" of AGN selection at optical/IR wavelength but opens up the possibility of studying the connection between nuclear black hole activity and host galaxy properties. We will come back more extensively in Sect. 3.2 to the issue of the overall decomposition of the AGN–galaxy spectral energy distribution in large multiwavelength surveys.

As for the general evolution of the optically selected QSO luminosity function, it has been known for a long time that luminous QSOs were much more common at high redshift ($z \sim 2$). Nevertheless, it is only with the aid of the aforementioned large and deep surveys covering a wide enough area of the distance-luminosity plane that it was possible to put sensible constraints on the character of the observed evolution. The most recent attempts (Croom et al. 2009) have shown unambiguously that optically selected AGN do not evolve according to a simple PLE, but instead more luminous objects peaked in their number densities at redshifts higher than lower luminosity objects, as shown in the left panel of *Fig. 11-14*.

We close this section with a brief discussion of the current status of IR AGN LF studies.

According to the AGN unification paradigm, obscuration comes from optically thick dust blocking the central engine along some lines of sight. The temperature in this structure, which

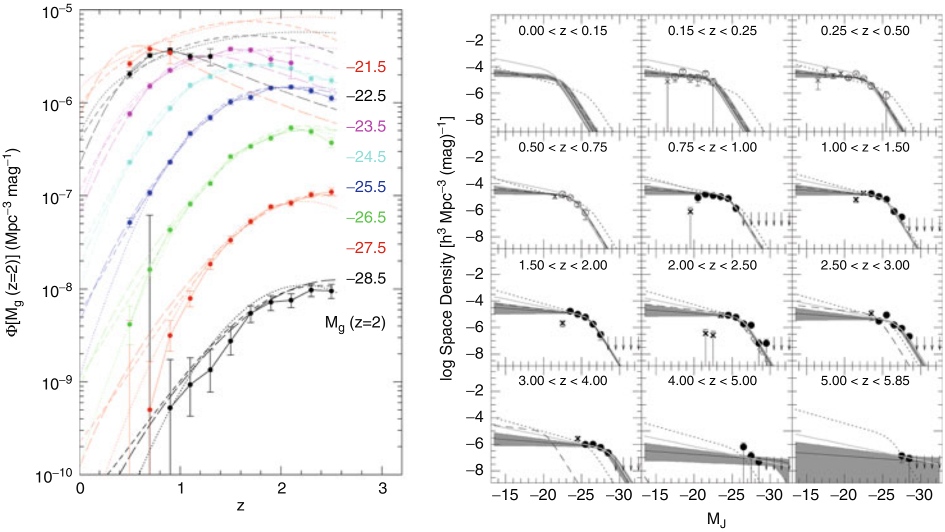


Fig. 11-14

***Left*: The combined 2SLAQ and SDSS optical QSO luminosity function plotted as a function of redshift for different absolute g-band magnitude intervals (the brightest at the *bottom* of the plot and the faintest at the *top*). The measured LF is compared to the best-fit PLE model (*dotted lines*), smooth LDDE model (*long dashed lines*), and LADE (Luminosity and density evolution) model (*short dashed lines*). Adopted from (Croom et al. 2009); *Right*: J-band luminosity function of mid-IR-selected AGN for several redshift bins. The *crosses* show points that were not used in the fits. The best-fit LADE, PLE, and pure PDE models are shown by the *solid*, *dashed*, and *dotted line*, respectively, although only the LADE model is an acceptable fit to the data. The *shaded area* shows the 2σ confidence region for the LADE fit. For reference, the *solid light gray line* shows the best-fit LADE model to a sample from a combined IR/X-ray selection (From Assef et al. (2011))**

can range up to 1,000 K (the typical dust sublimation temperature), and the roughly isotropic emission toward longer wavelengths should make both obscured and unobscured AGNs very bright in the mid- to far-infrared bands. This spectral shift of absorbed light to the IR has allowed sensitive mid-infrared observatories (*IRAS, ISO, Spitzer*) to deliver large numbers of AGN (see, e.g., Treister et al. 2006).

The problem with IR studies of AGN evolution, however, lies neither in the *efficiency* with which growing supermassive black holes can be found nor with the *completeness* of the AGN selection, which is clearly high and (almost) independent of nuclear obscuration but rather in the level of *contamination*. IR counts are, in fact, dominated by star-forming galaxies at all fluxes. This, and the lack of clear spectral signatures in the nuclear, AGN-powered emission in this band, implies that secure identification of AGN in any IR-selected catalog often necessitates additional information from other wavelengths, usually radio, X-rays, or optical spectroscopy.

Indeed, unlike the case of the CXRB, AGN contribute only a small fraction (up to 2–10%) of the cosmic IR background radiation (Treister et al. 2006), and similar fractions are estimated for the contribution of AGN at the "knee" of the total IR luminosity function at all redshifts.

Nonetheless, tremendous progress has been achieved in recent years, thanks to more refined color-selection criteria (Stern et al. 2005) that are little affected by contamination and provide reliable AGN samples, albeit with some well understood completeness biases against AGN that are faint with respect to their hosts and z ~ 4.5 type 1 AGN.

Thus, deep surveys with extensive multiwavelength coverage can still be used to track the evolution of active galaxies in the mid-infrared (see, e.g., Assef et al. 2011). Strengthening similar conclusions discussed above from other wavelengths, IR-selected AGN do not appear to evolve following either the PLE or PDE parametrizations but require significant differences in the evolution of bright and faint sources, with the number density of the former declining more steeply with decreasing redshift than that of the latter (see the right panel in ➲ *Fig. 11-14*).

2.1.4 Bolometric Luminosity Functions

We have seen in the previous sections how a qualitatively consistent picture of the main features of AGN evolution is emerging from the largest surveys of the sky in various energy bands. Strong (positive) redshift evolution of the overall number density as well as marked differential evolution (with more luminous sources being more dominant at higher redshift) characterizes the evolution of AGN.

Fundamental constraints on the physical evolution of the accretion-powered emission over cosmological times, like the ones we will discuss later in ➲ Sect. 3, require, ideally, a good knowledge of the *bolometric* luminosity function of AGN. This, in turn, demands a detailed assessment of selection biases and a robust estimation of the AGN spectral energy distribution (SED).

A thorough and detailed understanding of the AGN SED as a function of luminosity (and, possibly, of redshift, but see ➲ Sect. 1.1 above) could in principle allow us to compare and cross-correlate the information on the AGN evolution gathered in different bands. As for the accuracy of our knowledge of the bolometric correction, we refer the reader to the studies of Marconi et al. (2004), Richards et al. (2006), and Hopkins et al. (2007). All of them consistently demonstrate that a luminosity-dependent bolometric correction is required in order to match type I (unabsorbed) AGN luminosity functions obtained by selecting objects in different bands.

Summarizing the discussion of the previous sections, ➲ *Fig. 11-15* shows a compilation of luminosity functions observed at various wavelengths. The observed mismatch among the

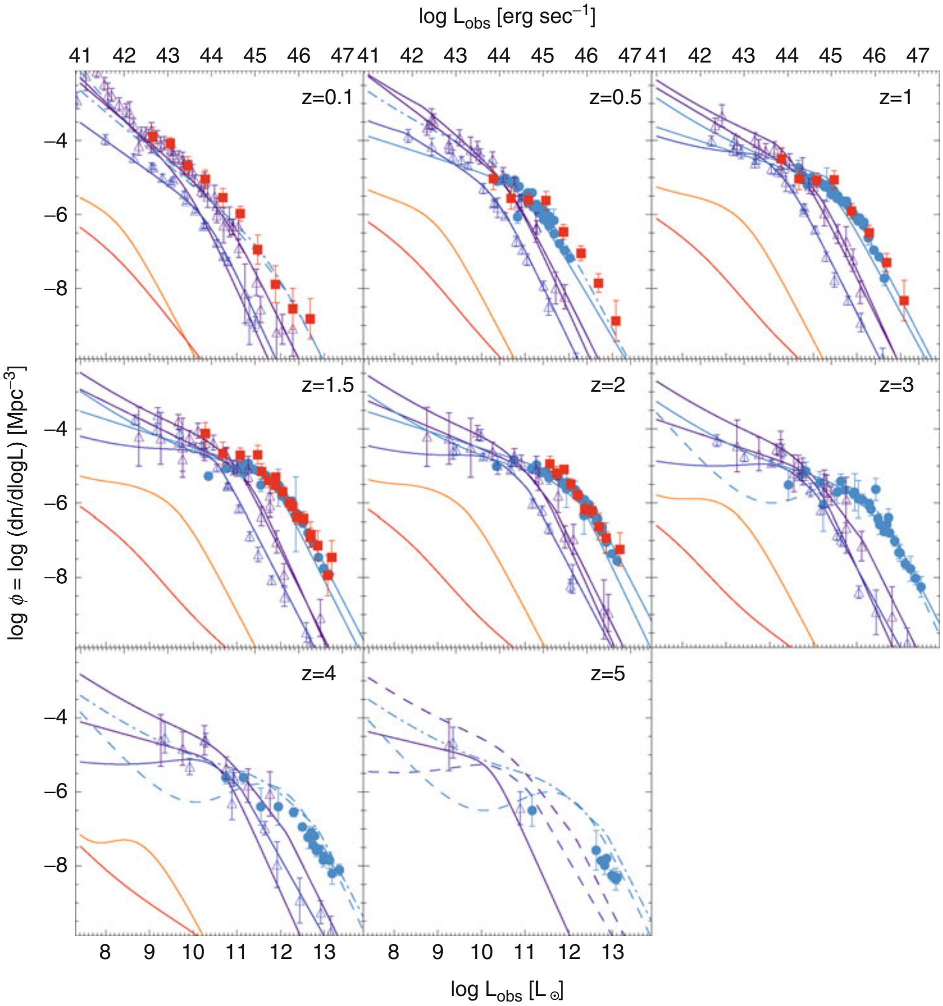


Fig. 11-15

A compilation of luminosity functions observed in various energy bands. The logarithm of the number of AGN per unit comoving volume and unit logarithm of luminosity is plotted as a function of the observed luminosity (in solar units). Observational points for IR (15 μm; *filled red squares*), B-band (*filled blue circles*), soft X-rays (0.5–2 keV; *empty blue triangles*), and hard X-rays (2–10 keV; empty *purple triangles*) are shown alongside published analytic fits for each band (*solid lines* in corresponding colors). The best-fit radio luminosity functions of steep- and flat-spectrum sources from Massardi et al. (2010) are also shown for comparison with *orange* and *red thick lines*, respectively. The observed mismatch among the various luminosity functions in *Fig. 11-15* is due to a combination of different bolometric corrections and incompleteness due to obscuration (Courtesy of P. Hopkins)

various LF observed at all redshift is due to a combination of different bolometric corrections and incompleteness due to obscuration. In fact, adopting a general form of luminosity-dependent bolometric correction, and with a relatively simple parametrization of the effect of the obscuration bias on the observed LF, Hopkins et al. (2007) were able to project the different observed luminosity functions in various bands into a single bolometric one, $\phi(L_{\rm bol})$ (➋ *Fig. 11-17*). As a corollary from such an exercise, we can then provide a simple figure of merit for AGN selection in various bands by measuring the bolometric energy density associated with AGN selected in that particular band as a function of redshift. We show this in the left panel of ➋ *Fig. 11-16* for four specific bands (hard X-rays, soft X-rays, UV, and mid-IR). From this, it is obvious that the reduced incidence of absorption in the 2–10-keV band makes the hard X-ray surveys recover a higher fraction of the accretion power generated in the universe than any other method.

While optical QSO surveys miss more than three quarters of all AGN of any given $L_{\rm bol}$, hard X-ray selection only fails to account for about one third (up to 50%) of all AGN, the most heavily obscured (Compton thick) ones, as shown in the right panel of ➋ *Fig. 11-16*. A common feature apparent from such a figure is that the effects of obscuration appear to be more severe at lower intrinsic luminosities, an observational fact that has been discussed previously in the context of X-ray surveys of AGN (see ➋ Sect. 1.1 above). It is important to note that the high missed fraction for mid-IR-selected AGN is a direct consequence of the need for (usually optical) AGN identification of the IR sources, so that optically obscured active nuclei are by and large missing in the IR AGN luminosity functions considered here.

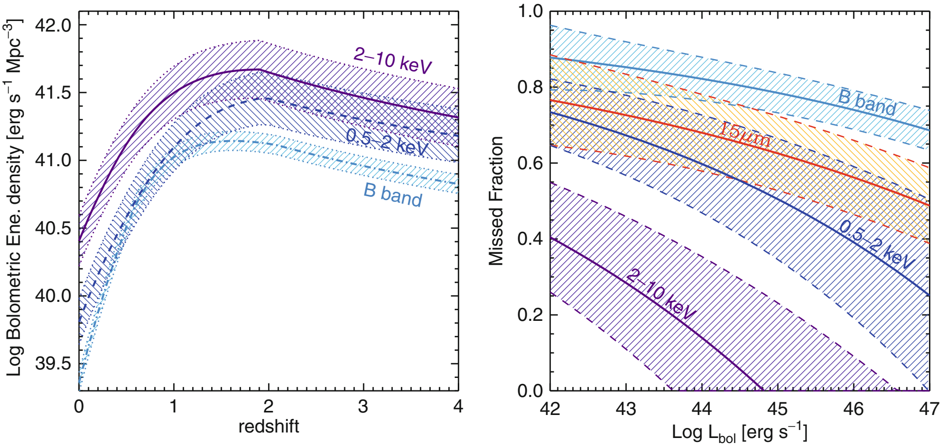


Fig. 11-16

***Left*: The redshift evolution of the bolometric energy density for AGN selected in different bands. Bolometric corrections from Hopkins et al. (2007) have been used, and the shaded areas represent the uncertainty coming from the bolometric corrections only. *Right*: The fraction of AGN missed by observations in any specific band as a function of the intrinsic bolometric luminosity of the AGN. *Red*, *light blue*, *dark blue*, and *purple-shaded* areas correspond to rest-frame mid-IR (15 μm), UV (B-band), soft X-rays (0.5–2 keV), and hard X-rays (2–10 keV), respectively. The uncertainty on the missed fractions depend on the uncertainties of the bolometric corrections and on the shape of the observed luminosity functions only**

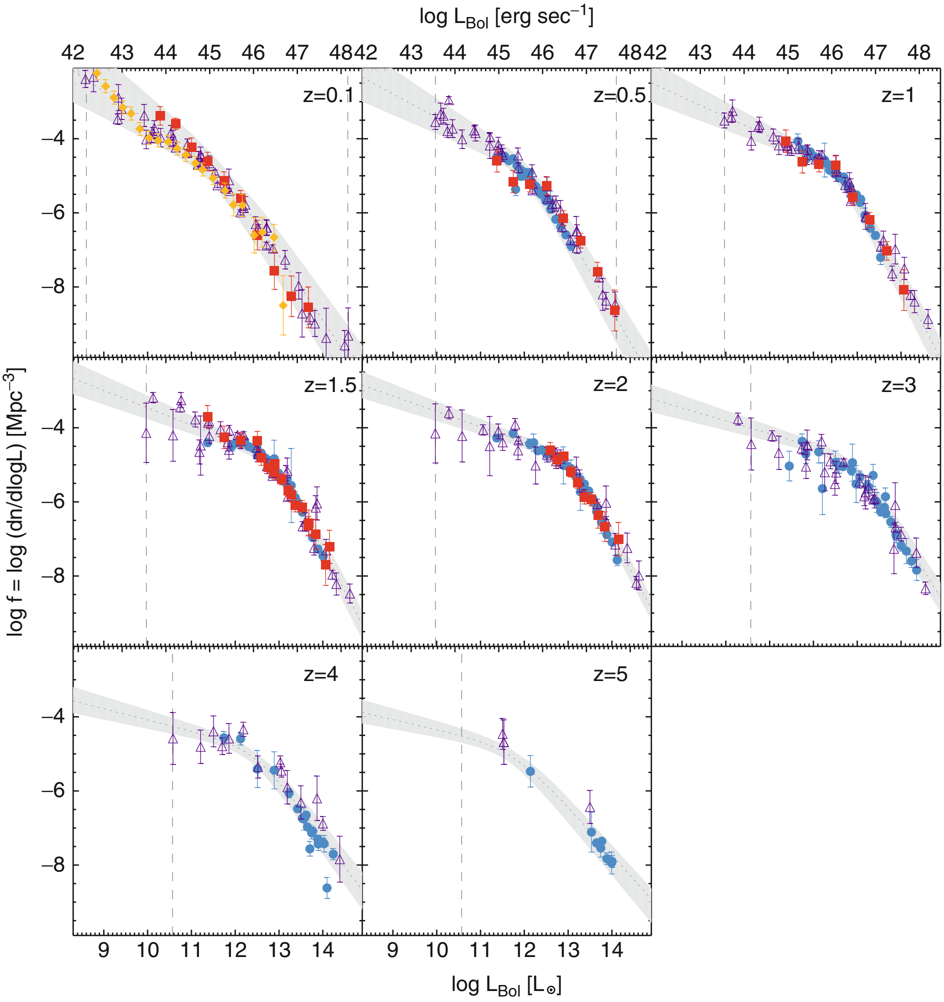
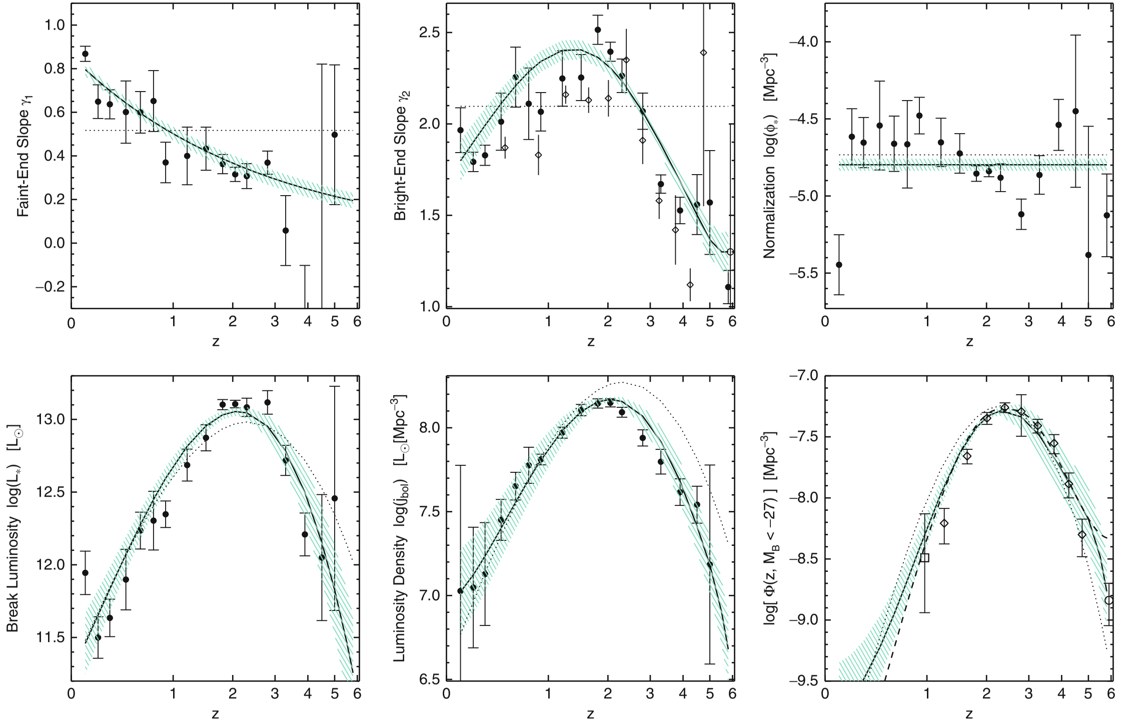


Fig. 11-17

Bolometric AGN luminosity function (*gray band*) as a function of redshift, as calculated by Hopkins et al. (2007). The different *symbols* and *colors* refer to different bands from which data have been extracted: *solid blue circles* are optical data, *filled red squares* IR (at 15 μm), *blue* and *purple triangle* are soft and hard X-ray data, respectively, and the *filled orange diamonds* are luminosities from emission lines. The *vertical dashed lines* bracket the observational limits. We refer the reader to Hopkins et al. (2007) for a more detailed description of the data and methodology used to extract the bolometric luminosity function (Data points courtesy of P. Hopkins)

Figure 11-18 shows the evolution of the parameters of the analytic fit to the bolometric LF data. They encompass our global knowledge of the evolution of accretion power onto nuclear black holes throughout the history of the universe. The three bottom panels reveal the overall



■ Fig. 11-18

Best-fit AGN bolometric LF double power-law parameters as a function of redshift. *Symbols* show the best-fit values to data at each redshift, *dotted lines* the best-fit PLE model, and *solid lines* the best-fit full model (a luminosity and density evolution one). Although PLE is appropriate for a lowest order fit, both the bright- and faint-end slopes evolve with redshift to high significance. The *bottom right panel* shows the predicted number density of bright optical quasars from the full fit (*solid line*), compared to that observed ones (From Hopkins et al. (2007))

increase in AGN activity with redshift, up to $z \approx 2$, and the mirroring high-redshift decline. At the center, the total integrated luminosity density evolution mark the epochs of rapid build up of the SMBH mass density. On the lower left, the evolution in the break luminosity $L_{\mathrm{bol},*}$ indicates that the "typical" accreting black holes were significantly more luminous at $z \approx 2$ than now, a different way of looking at AGN "downsizing." This is accompanied by a progressive steepening of the faint-end slope of the LF (upper left panel): Low-luminosity AGN become more and more dominant in the overall number density of AGN as time progresses.

Such a detailed view of the evolution of active galactic nuclei, with its distinctive signatures of "downsizing," has lent additional support to the notion that the lives of growing black holes must be intimately linked to those of their host galaxies. Indeed, both galaxies and black holes show signs of a similar differential evolution. The very term "downsizing" was first used by Cowie et al. (1996) to describe the finding that actively star-forming galaxies at low redshift have smaller masses than actively star-forming galaxies at $z \sim 1$. It has come to identify, in the current cosmology jargon, a variety of possibly distinct phenomena, not just related to the epoch of star formation, but also to that of star-formation quenching or galaxy assembly (see Chap. 10). Our current understanding of AGN evolution, encapsulated in the observable evolution of their bolometric luminosity function, emphatically suggests that growing nuclear black holes take part in this global process of structure formation.

2.2 AGN Clustering and the Large-Scale Structure of the Universe

The bolometric luminosity function of AGN provides the basic tool to describe the differential evolution of growing black holes. Some key properties of AGN, however, remain impossible to determine on the sole basis of the observed LF. As it is increasingly difficult to measure black hole masses at high redshift, and currently only possible for bright, unobscured broad line QSOs, we do not have robust, direct, observational constraints on the distribution of AGN Eddington ratios beyond the local universe. The Eddington ratio distribution, in turn, depends on the details of the average AGN lightcurves which could reveal important details of the physical processes driving the accreting gas toward the black hole.

The spatial distribution of QSOs (clustering) in the sky could provide such an alternative method to estimate lifetimes (Martini and Weinberg 2001; see also Chap. 8). In the current ΛCDM paradigm for structure formation, more clustered objects are rarer and live in more massive dark matter structures (or "halos"). Thus, if AGN are strongly clustered, their hosts must be rare objects, too, and the effective AGN lifetime must be long, in order for such a rare "parent" population to account for the total AGN luminosity density observed. If, on the other hand, their clustering is comparable to the clustering of smaller, less massive, dark matter halos, their host are more common and their luminous phases must therefore have short duration.

A commonly used technique for measuring the spatial clustering of a class of objects is the two-point correlation function $\xi(r)$, which measures the excess probability dP above a random distribution of finding an object in a volume element dV at a distance r from another randomly chosen object:

$$dP = n[1 + \xi(r)]dV. \tag{11.4}$$

where n is the mean number density of objects. In the scale range between a few tens of kpc and a few tens of Mpc, for most classes of astronomical objects $\xi(r)$ can be described by a single power law:

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma}. \tag{11.5}$$

where r_0 is the *correlation, or clustering, length*, defined as the scale at which the two-point correlation function is equal to unity.

Unfortunately, a direct comparison of the measured clustering length of AGN with that expected for dark matter halos of different masses is hampered by the fact that, according to current theories of structure formation, galaxies (and their nuclear black holes) do not follow the distribution of the underlying matter but form in the high-density peaks of the dark matter field. The *bias* of any astrophysical population X is defined as the (square root of the) ratio between the two-point autocorrelation functions of population X and of the dark matter (DM) halos: $b_{\mathrm{X,DM}}(r) \equiv \sqrt{\xi_{\mathrm{X}}(r)/\xi_{\mathrm{DM}}(r)}$.

Many groups have now been able to measure the clustering of AGN at different luminosities, bands, scales, and redshifts (see, e.g., the recent review of Shankar 2009 for a complete list of references). Overall, the clustering length of quasars appears to be an increasing function of redshift but does not depend strongly on luminosity.

As shown in ❯ *Fig. 11-19*, the bias of optically selected (broad line) AGN increases with redshift following an evolution at approximately constant dark matter halo mass (since halos of a fixed mass are progressively more clustered toward higher redshift), in the range $\log M_{\mathrm{DM}} \simeq 12.5\text{–}13\,\mathrm{h}^{-1} M_{\odot}$ at redshifts $z < 3.5$.

Instead, X-ray selected objects (Allevato et al. 2011), both obscured and unobscured, reside in more massive DM structures at all redshifts $z < 2.25$, with a typical mass of the hosting halos constant over time in the range $\log M_{\mathrm{DM}} \simeq 13\text{–}13.5\,\mathrm{h}^{-1} M_{\odot}$.

By combining the number density of AGN with that of the hosting dark matter halos, one can estimate an AGN duty cycle and a corresponding average lifetime. The observed biases of the rare, luminous broad-lined quasars imply timescales of the order of 10^7–10^8 years, increasing with redshift, as the massive halos typically hosting AGN become increasingly rare.

For X-ray selected AGN, a larger duty cycle is inferred, which translates into an AGN lifetime of ~ 0.1–1 Gyr, about one order of magnitude longer than that estimated for optically bright QSOs at the same redshift. This is mainly due to the higher number density and higher bias of AGN found in X-ray selected samples.

Numerical simulations of merger-induced AGN activity in a cosmological context have shown (Bonoli et al. 2009) that the clustering of optically selected quasars is well explained by a model in which these objects are triggered by major merger events. The difference between optically selected and the (lower luminosity) X-ray selected AGN of ❯ *Fig. 11-19* might suggest that X-ray selected AGN are triggered by different (secular) processes which may be capable of fueling luminous AGN in the gas-rich environment of star-forming galaxies at high redshift. The same models also predict an increase in AGN duty cycle for the brightest quasars at halo masses larger than $10^{12} M_{\odot}$ but fail to reproduce the large biases for less luminous X-ray selected AGN at $z \sim 1$, possibly pointing (again) toward the need for a larger variety of AGN triggering mechanisms for this class of objects.

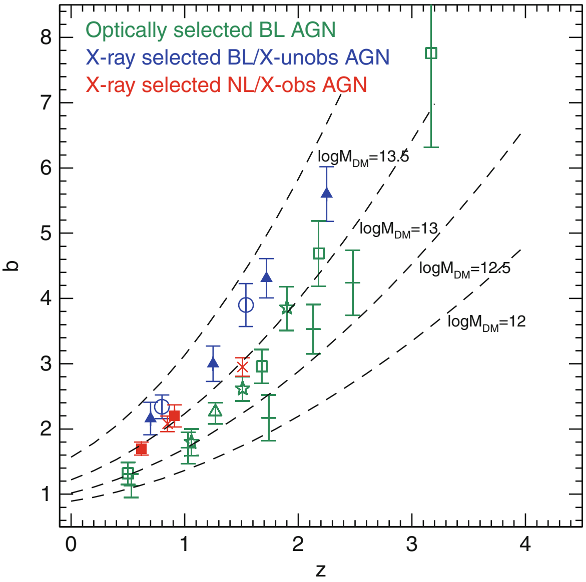


Fig. 11-19

Bias parameter as a function of redshift for various AGN surveys. The range of the bolometric luminosity probed is given in parenthesis: (i) optically selected BL AGN: *green-crosses* (2dF; $45.3 < \log L_{\mathrm{bol}} < 46.7$), *green-stars* (2dF; $45.5 < \log L_{\mathrm{bol}} < 47.4$), *green-open squares* (SDSS; $45.6 < \log L_{\mathrm{bol}} < 46.9$), and *green-open triangles* (SDSS; $\log L_{\mathrm{bol}} \sim 46.5$); (ii) X-ray selected unobscured AGN: *blue triangles* and *blue open circles* (XMM-COSMOS; $45 < \log L_{\mathrm{bol}} < 45.7$); and (iii) X-ray selected obscured AGN: *red squares* and *red crosses* (XMM-COSMOS; $44.1 < \log L_{\mathrm{bol}} < 44.6$). See Allevato et al. (2011) for a similar figure and the full list of references for the observed data points. The *dashed lines* show the expected bias evolution of typical DM halo masses

3 Cosmology I: The Growth of Supermassive Black Holes in Galaxies

As we have discussed in the previous section, the strong cosmological evolution of the quasar population was recognized early on by observers in essentially all bands of the electromagnetic spectrum. In the early 1990s, deep optical surveys of star-forming galaxies began to probe the cosmological evolution of the rate at which stars are formed within galaxies, thus providing robust constraints for models of galaxy formation and evolution (the so-called Lilly–Madau plot; Madau et al. 1996). It was soon clear that QSOs luminosity density and star formation rate (SFR) density evolved in similar fashion, being much higher in the past, with a possible (very broad) peak at $z \approx 2$ (Boyle and Terlevich 1998).

In the previous section, we have traced the history of the study of AGN luminosity functions in various spectral bands, closing with an assessment of our current understanding of the bolometric luminosity function evolution. A reliable census of the bolometric energy output of growing supermassive black holes (see, e.g., the central bottom panel of *Fig. 11-18*) allows a more direct estimate of the global rate of mass assembly in AGN and an interesting comparison with that of stars in galaxies. Together with the tighter constraints on the "relic" SMBH mass density in the local universe, $\rho_{\mathrm{BH},0}$, provided by careful application of the scaling relations between black hole masses and host spheroids, this enables meaningful tests of the classical "Soltan argument" (Soltan 1982), according to which the local mass budget of black holes in galactic nuclei should be accounted for by integrating the overall energy density released by AGN, with an appropriate mass-to-energy conversion efficiency.

Many authors have carried out such a calculation, either using the CXRB as a "bolometer" to derive the total energy density released by the accretion process (Fabian and Iwasawa 1999) or by considering evolving AGN luminosity functions (Yu and Tremaine 2002; Marconi et al. 2004; Merloni and Heinz 2008). Despite some tension among the published results that can be traced back to the particular choice of AGN LF and/or scaling relation assumed to derive the local mass density, it is fair to say that this approach represents a major success of the standard paradigm of accreting black holes as AGN power sources, as the radiative efficiencies needed to explain the relic population are within the range ≈ 0.06–0.20, predicted by standard relativistic accretion disc theory (Novikov and Thorne 1973).

In this section, we begin with a schematic account of the current constraints on the black hole mass density growth and discuss some recent attempts to compare it on a quantitative level with the observed growth of the galaxy population. This will be followed by the (related) discussion of the possible evolution of the scaling relations.

3.1 A Global View of the Accretion History of the Universe

Under the standard assumption that black holes grow mainly by accretion, their cosmic evolution can be calculated from the bolometric luminosity function of AGN $\phi(L_{\mathrm{bol}}, z)$, where $L_{\mathrm{bol}} = \epsilon_{\mathrm{rad}} \dot{M} c^2$ is the bolometric luminosity produced by an SMBH accreting at a rate of $\dot{M}$ with a *radiative* efficiency ϵ_{rad}. The non-negligible fraction of the AGN population which is unaccounted for in current surveys, the so-called Compton-thick AGN (see ➋ Sect. 2.1.2 above), is usually included in the bolometric luminosity function by assuming a redshift-invariant column density distribution as measured in the very local universe and an overall number density of heavily obscured AGN that fits the CXRB.

The total, integrated mass density in supermassive black holes can then be computed as a function of redshift:

$$\frac{\rho_{\mathrm{BH}}(z)}{\rho_{\mathrm{BH},0}} = 1 - \int_0^z \frac{\Psi_{\mathrm{BH}}(z')}{\rho_{\mathrm{BH},0}} \frac{dt}{dz'} dz', \tag{11.6}$$

where the black hole accretion rate (BHAR) density is given by

$$\Psi_{\mathrm{BH}}(z) = \int_0^\infty \frac{(1-\epsilon_{\mathrm{rad}})L_{\mathrm{bol}}}{\epsilon_{\mathrm{rad}} c^2} \phi(L_{\mathrm{bol}}, z) dL_{\mathrm{bol.}} \tag{11.7}$$

and

$$\frac{dt}{dz} = -\left[(1+z)H_0\sqrt{(1+z)^3\Omega_m + \Omega_\Lambda}\right]^{-1}. \tag{11.8}$$

The exact shape of $\rho_{\mathrm{BH}}(z)$ and $\Psi_{\mathrm{BH}}(z)$ then depends only on the local black hole mass density $\rho_{\mathrm{BH},0}$ and the (average) radiative efficiency ϵ_{rad}.

We can then link the growth of SMBH from (11.6) to the growth of stellar mass in galaxies. To do so, we will use the Hopkins et al. (2007) bolometric LF of AGN (see Sect. 2.1.4 above).[4] Because local SMBH are observed to correlate with spheroids only, we introduce the parameter $\lambda(z)$, the ratio of the mass in discs, and irregulars to that in spheroids at any redshift, so that the total stellar mass density can be expressed as $\rho_*(z) = \rho_{\rm sph}(z) + \rho_{\rm disk+irr}(z) = \rho_{\rm sph}(z)[1+\lambda(z)]$.

We can now assume that $\lambda(z)$ evolves according to $\lambda(z) = \lambda_0(1+z)^{-\beta}$, where λ_0 is the value of the disc to spheroid mass density ratio in the local universe. Also, we assume that the mass density of spheroids and supermassive black holes evolve in parallel, modulo a factor $(1+z)^{-\alpha}$, obtaining a prediction for the observable stellar mass density evolution as traced by SMBH growth:

$$\rho_*(z) = \mathcal{A}_0 \rho_{\rm BH}(\epsilon_{\rm rad}, z)(1+z)^{-\alpha}[1+\lambda_0(1+z)^{-\beta}]. \tag{11.9}$$

where $\mathcal{A}_0$ is the constant of proportionality in the SMBH mass–spheroid mass relation. By taking the derivative of (11.9), accounting for stellar mass loss, an expression is also found for the corresponding star formation rate (SFR) density evolution:

$$d\rho_*(z)/dt = \Psi_*(z) - \int_{z_i}^{z} \Psi_*(z') \frac{d\chi[\Delta t(z'-z)]}{dt} \frac{dt}{dz'} dz', \tag{11.10}$$

where $\chi[\Delta t(z'-z)]$ is the fractional mass loss that a simple stellar population experiences after a time Δt (corresponding to the redshift interval $(z'-z)$), and z_i is the redshift of (instantaneous) formation of the first stellar populations.[5]

With these expressions, we fit observational data points of both $\rho_*(z)$ and SFR(z). For each choice of $\rho_{\rm BH,0}$, λ_0, and of the critical accretion rate $\dot{m}_{\rm cr}$, the fitting functions depend only on three parameters: α, β, and the radiative efficiency $\epsilon_{\rm rad}$. One example of such fits is shown in *Fig. 11-20* for the specific case $\rho_{\rm BH,0} = 4.2 \times 10^5 M_\odot\,{\rm Mpc}^{-3}$ (Shankar 2009), and $\lambda_0 = 0.3$ (Fukugita and Peebles 2004).

Because the drop in the AGN integrated luminosity density at low z is apparently faster than that in the SFR density, the average black hole to spheroid mass ratio must evolve (slightly) with lookback time ($\alpha > 0$). This result is independent of the local black hole mass density and independent of λ_0. For the particular example shown here, the average radiative efficiency turns out to be $\epsilon_{\rm rad} = 0.08^{+0.01}_{-0.02}$, while we obtain $\alpha = 0.35^{+0.22}_{-0.3}$ (both shown with 3-σ confidence bounds). At face value, this would imply a very mild evolution of the average $M_{\rm BH}/M_{\rm sph}$ mass ratio. We will discuss in Sect. 3.2.1 how these constraints compare with recent efforts to directly measure the ratio of black hole to host galaxy mass at high redshift.

This simple exercise should make clear that the available constraints on SMBH growth from the observed bolometric LF are robust enough to provide interesting nontrivial insight into the cosmological coevolution of AGN and galaxies.

3.1.1 The Evolution of the SMBH Mass Function

Despite the relative successes of "Soltan argument" – like calculations of the integral evolution of the SMBH mass density, it is obvious that a much greater amount of information is contained

[4] As discussed in Marconi et al. (2004), in order to correctly estimate the total bolometric output of an AGN, care should be taken in avoiding double counting of the IR reprocessed emission. This appears not have been done in Hopkins et al. (2007), so we correct the bolometric luminosities by 30% to account for this.

[5] An analogous term for $\rho_{\rm BH}$, due to the ejection of SMBHs from galaxy halos after a merger event, is much more difficult to estimate and is neglected here.

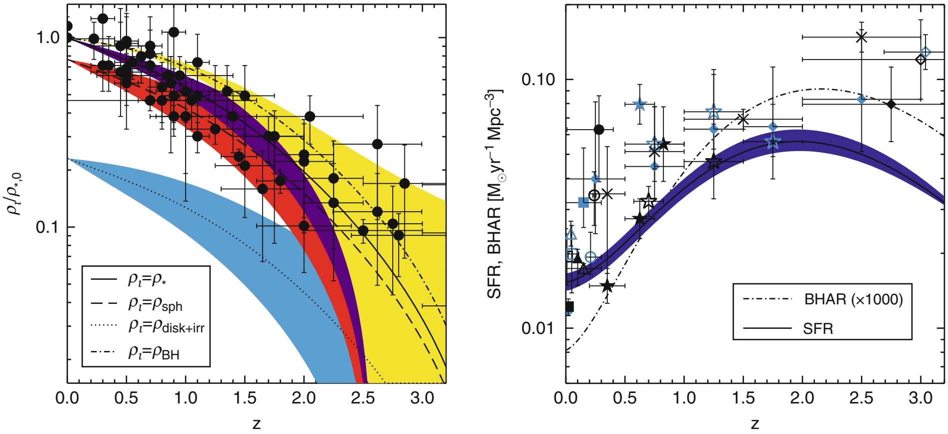


Fig. 11-20

Left: Evolution of the stellar mass density as a function of redshift (_black points_, observations), where the density is given as a ratio to the local value, $\rho_{*,0} = 5.6 \times 10^8 M_\odot$ Mpc^{-3} (Cole et al. 2001). _Shaded areas_ represent 1-sigma confidence intervals of the model fits. _Solid black line_ with purple-shaded area shows the best joint fit from (11.9) and (11.10) to both stellar mass and SFR density points (_left_ and _right panels_), while the _dot-dashed line_ with _yellow-shaded area_ marks the normalized evolution of the SMBH mass density only. The slight offset between the two is compensated by a change in the normalization of the average black hole to spheroid mass ratio with redshift (see text for details). The _dashed_ (_dotted_) _line_ with _red (blue)-shaded_ area shows the relative growth of the mass density in spheroids (discs). Values of $\lambda_0 = 0.3$ (Fukugita and Peebles 2004) and $\rho_{\mathrm{BH},0} = 4.2 \times 10^5 M_\odot$ Mpc^{-3} (Shankar 2009) are adopted here. _Right_: The corresponding best-fit relation for the SFR density evolution, from (11.10) is shown with a _solid line_ and _dark-blue-shaded_ area. The _dash-triple-dotted line_ is (1,000 times) the black hole accretion rate density $\Psi_{\mathrm{BH}}(z)$ (BHAR). It appears that the BHAR declines slightly faster than the SFR, another way to emphasize the need of an evolution in the average $M_{\mathrm{BH}}/M_{\mathrm{sph}}$ ratio

in the _differential_ distributions (mass and luminosity functions). We will now discuss attempts to use this information to constrain the evolution of the mass function of SMBH.

As opposed to the case of galaxies, where the direct relationship between the evolving mass functions of the various morphological types and the distribution of star-forming galaxies is not straightforward due to the never-ending morphological and photometric transformation of the different populations, the situation in the case of SMBH is much simpler. For the latter case, we can assume their evolution is governed by a continuity equation (Merloni and Heinz 2008, and references therein), where the mass function of SMBH at any given time can be used to predict the mass function at any other time, provided the distribution of accretion rates as a function of black hole mass is known. Such an equation can be written as

$$\frac{\partial\psi(\mu,t)}{\partial t} + \frac{\partial}{\partial\mu}\left(\psi(\mu,t)\langle\dot{M}(\mu,t)\rangle\right) = 0. \tag{11.11}$$

where $\mu = \log M$ is the black hole mass in solar units and $\psi(\mu,t)$ is the SMBH mass function at time t. $\langle\dot{M}(\mu,t)\rangle$ is the average accretion rate of SMBH of mass M at time t and can be defined through a "fueling" function, $F(\dot{\mu}, \mu, t)$, which describes the distribution of accretion rates for objects of mass M at time t

$$\langle \dot{M}(\mu, t) \rangle = \int \dot{M} F(\dot{\mu}, \mu, t)\, \mathrm{d}\dot{\mu}, \tag{11.12}$$

Such a fueling function is not a priori known, and observational determinations thereof have been possible in any robust sense only for the extremes of the overall population. However, the AGN fueling function can be derived by inverting the integral equation that relates the luminosity function of the population in question with its mass function. And so, we can write

$$\phi(\ell, t) = \int F(\ell - \zeta, \mu, t)\psi(\mu, t)\, \mathrm{d}\mu \tag{11.13}$$

with the definitions $\ell \equiv \log L_{\mathrm{bol}}$ and $\zeta \equiv \log(\epsilon_{\mathrm{rad}} c^2)$, with ϵ_{rad} the radiative efficiency, here assumed to be constant.

Using this approach, (❯ 11.11) can be integrated backward from $z = 0$, where we have simultaneous knowledge of both the mass function, $\psi(\mu)$, and the luminosity function, $\phi(\ell)$, thus evolving the SMBH mass function *backward* in time, up to where (i) reliable estimates of the AGN luminosity functions are available and (ii) the accumulated error in the mass function becomes of the order of the mass function itself.

The first thing to notice from such an approach is that the different shapes of the observed SMBH mass function $\psi(\mu)$ (that decays exponentially at high masses) and AGN LF $\phi(\ell)$ (well described by a double power law) necessitate a broad distribution of accretion rate (Merloni and Heinz 2008): AGN, as a population, cannot be simply characterized by an on–off switch at fixed Eddington ratio. Instead, integration of (❯ 11.11) gives insight on the relative importance of massive black hole growth at different accretion rates.

❯ *Figure 11-21* shows the number density evolution as a function of redshift for black holes of constant mass (in the range $8 < \log(M_{\mathrm{BH}}/M_\odot) < 8.5$) at different Eddington ratios. Numerically, the AGN population is always dominated by slowly accreting objects, but the observed flattening of the bolometric LF shape (see ❯ *Fig. 11-18*) implies that the relative number of rapidly accreting black holes increases significantly with redshift. In terms of grown mass, however, high-Eddington-ratio AGN strongly dominate the budget, as shown in the top panel of ❯ *Fig. 11-21*: most of the mass of a typical $\approx 10^8 M_\odot$ black hole has been accumulated in (short-lived) episodes of rapid accretion, between a few and a few tens of percent of the Eddington luminosity.

The specific instantaneous ratio of black hole mass to accretion rate as a function of SMBH mass defines a timescale, the so-called *growth time*, or mass doubling time. The redshift evolution of the growth time distribution can be used to identify the epochs when black holes of different sizes grew the largest fraction of their mass: Black holes with growth times longer than the age of the universe are not experiencing a major growth phase, which must have necessarily happened at earlier times.

❯ *Figure 11-22* (left) shows that, according to this simple estimate, while at $z < 1$ only black holes with masses smaller than $10^7 M_\odot$ are experiencing significant growth, as we approach the peak of the black hole accretion rate density ($z \sim 1.5 - 2$), we witness the rapid growth of the *entire* SMBH population.

Solutions of the continuity equation also allow one to trace the growth of black holes of a given final (i.e., at $z = 0$) mass. The right-hand side panel of ❯ *Fig. 11-22* shows that, for the most massive black holes ($>10^9 M_\odot$), half of the mass was already in place at $z \sim 2$, while those with $M(z = 0) < 10^8 M_\odot$ had to wait until $z \sim 1$ to accumulate the same fraction of their final mass.

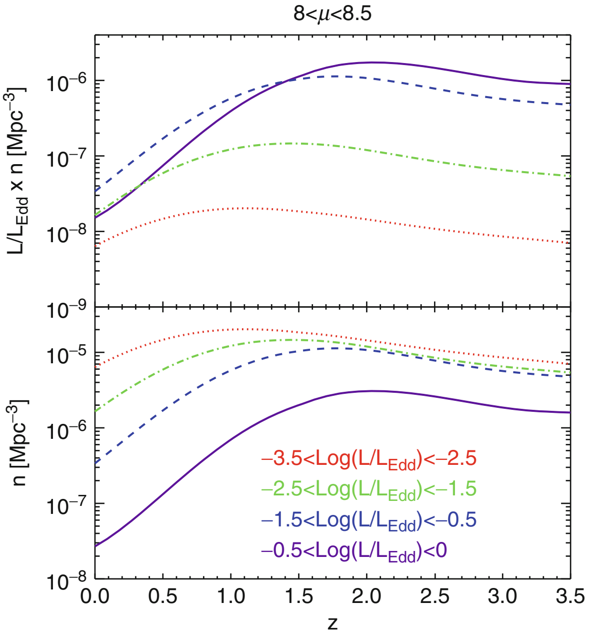


Fig. 11-21
***Bottom*: Evolution of the number density (objects per comoving Mpc) as a function of redshift for black holes of constant mass (in the range $8 < \log(M_{BH}/M_\odot) < 8.5$) at different Eddington ratios. *Solid* (*purple*), *dashed* (*blue*), *dot-dashed* (*green*), and *dotted* (*red*) *lines* correspond to intervals of Eddington ratio ranging from 1 to $10^{-3.5}$. *Top*: The product of average Eddington ratio times number density versus redshift. Despite being numerically sub-dominant, rapidly accreting black holes (i.e., those with $L/L_{Edd} > 3\%$) clearly dominate the mass assembly of SMBH in this range of masses**

3.2 The AGN-Galaxy Connection

The very existence of scaling relations between black holes and their host galaxies and the broad accretion rate distributions of AGN derived from the continuity equation approach imply that, as observed throughout the electromagnetic spectrum, growing black holes will display a large range of "contrast" with the host galaxy light.

The most luminous QSO, accreting at the highest Eddington ratios, will be able to outshine the stellar light from the galaxy, while less luminous, Seyfert-like AGN will have a global SED with a non-negligible contribution from the host (see also (➋ 11.3) above). At high redshift, when it becomes increasingly difficult to spatially separate the nuclear emission, unbiased AGN samples will have optical–NIR colors spanning a large range of intermediate possibilities.

➋ *Figure 11-23* nicely illustrates this point. It is taken from the analysis of an X-ray selected sample of AGN in the COSMOS field (Brusa et al. 2010), the largest fully identified and redshift complete AGN sample to date. It displays the slope of the rest-frame SED in the optical (α_{OPT}, between 0.3 and 1 μm) and NIR (α_{NIR} between 1 and 3 μm). Pure QSOs, i.e., objects in which the overall SED is dominated by the nuclear (AGN) emission, have a typical dip in the NIR region and would lie close to the empty blue star in the lower right corner (positive optical

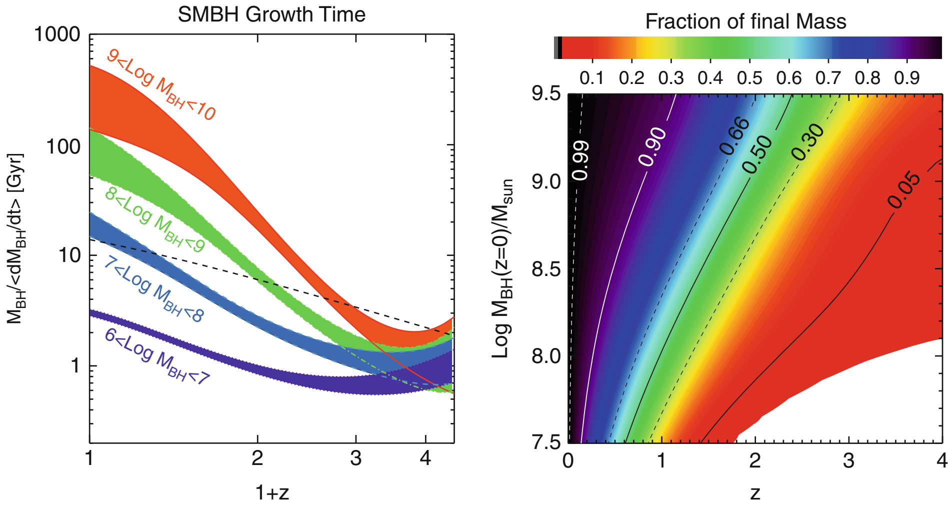


Fig. 11-22

***Left*: Average growth time of supermassive black holes (in Gyrs) as a function of redshift for different black hole mass ranges. The *dashed line* marks the age of the universe; only black holes with instantaneous growth time smaller than the age of the universe at any particular redshift can be said to be effectively growing. *Right*: The fraction of the final black hole mass accumulated as a function of redshift and final (i.e., at z = 0) mass is plotted as contours**

slope and negative NIR slope). The location of the X-ray selected AGN in *Fig. 11-23* shows instead that, in order to describe the bulk of the population, one needs to consider both the effects of obscuration (moving each pure QSO in the direction of the orange arrow) and an increasing contribution from stellar galaxy light (moving the objects toward the black stars in the upper part of the diagram).

This demonstrates that current multiwavelength extragalactic surveys are sensitive enough to disentangle the complex interplay between nuclear and galaxy light in the SED of more typical AGN. It is no coincidence that such surveys are beginning to probe the details of the coevolution of black holes and host galaxies on an object-by-object basis. In the following section, we will briefly discuss how one can use such information to observationally trace the evolution of the scaling relations between nuclear SMBH and their host galaxies.

3.2.1 Redshift Evolution of the Scaling Relations

Local scaling relations between black hole mass and structural properties of their (spheroidal) hosts have been unable to unambiguously determine the physical nature of the SMBH–galaxy coupling. A large number of theoretical models for the AGN–galaxy interaction responsible for establishing, for example, the $M - \sigma_*$ relation, have been proposed, all tuned to reproduce the $z = 0$ observations. One obvious way out of this impasse is the study of their *evolution*.

In recent years, a number of groups have employed different techniques to detect signs of evolution in any of the locally observed scaling relations. Only type 1 AGN, with unobscured

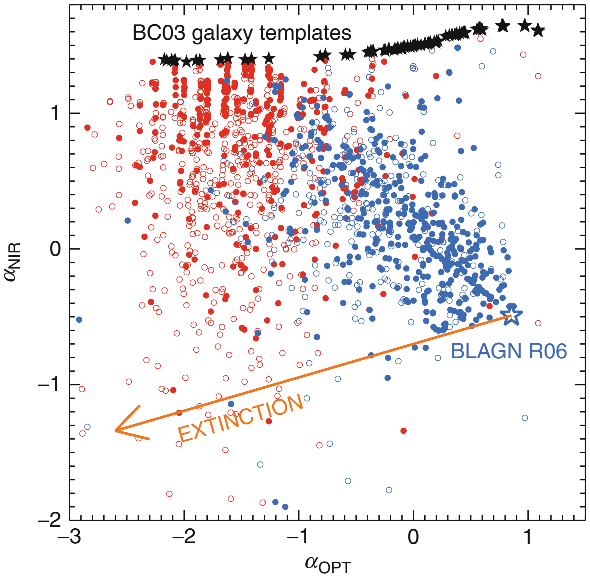


Fig. 11-23
Observed rest-frame SED slopes in the optical (α_{OPT}, between 0.3 and 1 μm) and NIR (α_{NIR} between 1 and 3 μm) for all (~1,650) X-ray selected AGN in the COSMOS survey. *Blue-filled circles* denote spectroscopically confirmed type 1 (*broad lined*) AGN, and *blue empty circles* denote candidate type 1 AGN from the photo-z sample. *Red-filled circles* are spectroscopically confirmed type 2 (*narrow lined*) AGN, empty *red circles* are candidate type 2 AGN from the photo-z sample. The *empty blue* star marks the colors of a pure intrinsic type 1 quasar SED (from Richards et al. 2006), while *black stars* are the loci of synthetic spectral templates of galaxies, with increasing levels of star formation from the left to the right. Nuclear obscuration, parametrized with a Calzetti extinction law, moves every pure type 1 AGN along the direction of the *orange arrow*

broad line region allow a simple direct estimate of BH masses, via the so-called "virial" or empirically calibrated "photoionization" method (Peterson et al. 2004). Based on existing samples of broad line QSOs, most efforts have been devoted to the study of the $M_{BH}-\sigma_*$ relation. For example, Salviander et al. (2007) have used narrow nebular emission lines ([OIII], [OII]) excited by the AGN emission in the nuclear region of galaxies as proxies for the central velocity dispersion and compared these to the black hole mass estimated from the broad line width of QSOs from $z \sim 0$ to $z \sim 1$. In this case, a large scatter has been found in the relation between M_{BH} and σ_*.

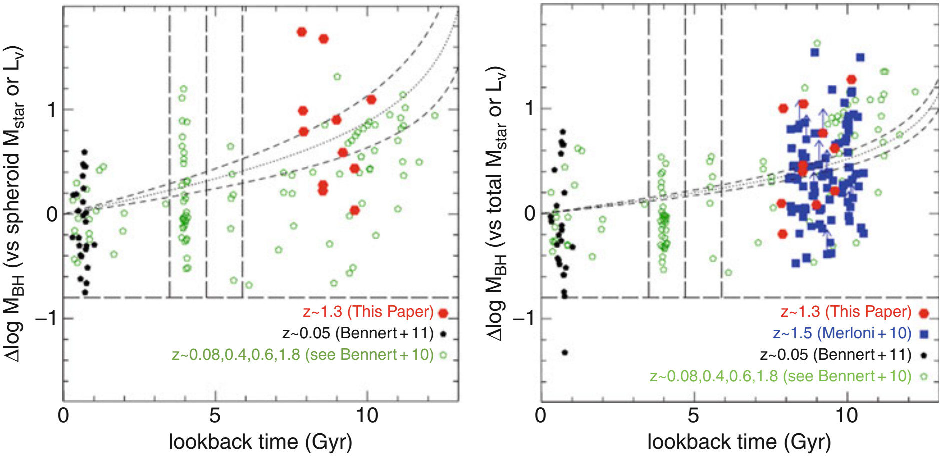
An alternative path is to study carefully selected samples of moderately bright AGN in narrow redshift ranges, where the host's stellar velocity dispersion can be measured directly from the absorption lines in high signal-to-noise spectra. These studies also found evidence of (strong) positive evolution of the M_{BH} to σ_* ratio compared to the local value (see Bennert et al. 2011, and references therein). This method, although promising and reliable, is quite inefficient and telescope-time consuming: Secure detection of spectral absorption features in massive ellipticals at $1 < z < 2$ require hundreds of hours of integration time on an 8-m class telescope.

When a good sampling of the AGN SED is instead available, rather than high-resolution, high signal-to-noise spectra, it is possible to try to decompose the overall spectral energy distribution into a nuclear and a galaxy component and derive in this way the physical properties of the host galaxies of unobscured AGN whose SMBH masses can be estimated from their broad lines (Merloni et al. 2010).

Other groups have chosen to try to derive information on the host mass of broad line AGN using multicolor image decomposition techniques. Due to the severe surface brightness dimming effects, employing these techniques for high-redshift QSOs becomes increasingly challenging, unless gravitationally lensed QSOs are selected. In all cases, very deep, high-resolution optical images (*HST*) are necessary to reliably disentangle the nuclear from the host galaxy emission.

The main result of these various investigations is that our estimates of the type 1 AGN host physical parameters are *inconsistent* with the hypothesis that they lie on the $z = 0$ scaling relation (see ❯ *Fig. 11-24*). At high redshift, bigger black holes are hosted in galaxies of a given mass as compared to what we observe locally. The best linear fit to the ensemble of observations shown in ❯ *Fig. 11-24* is $M_{\rm BH}/M_{\rm sph} \propto (1+z)^{2.1\pm0.3}$ for the black-hole-to-spheroid mass ratio and $M_{\rm BH}/M_{\rm host} \propto (1+z)^{1.41\pm0.12}$ for the black-hole-to-total host stellar mass ratio.

However, the objects for which this study can be made are selected essentially on the basis of the nuclear (AGN) luminosity and on the detectability of broad emission lines, clearly



■ Fig. 11-24

***Left*: Offset in log $M_{\rm BH}$ as a function of constant spheroid host galaxy mass (*red-filled pentagons*) with respect to the fiducial local relation of AGNs (*black-filled circles*). The offset as a function of constant stellar spheroid luminosity is overplotted (*green-open symbols*), corresponding to AGNs at different redshifts. The best linear fit derived here is overplotted as *dotted line* $M_{\rm BH}/M_{\rm sph} \propto (1+z)^{2.1\pm0.3}$; *dashed lines*: 1$\sigma$ range. *Right*: The same as in the *left panel* as an offset in log $M_{\rm BH}$ as a function of constant total host galaxy mass. The lines correspond to $M_{\rm BH}/M_{\rm host} \propto (1+z)^{1.41\pm0.12}$ (From Bennert et al. (2011), where a comprehensive list of references for the observational data points can also be found)**

leading to a bias toward more massive black holes, similar to Malmquist bias for luminosity-selected samples of standard candles. Can such a bias be responsible for the observed trends?

Let us consider in detail the effects on the observed systems of a given intrinsic scatter, σ_μ, in the $M_{\rm BH} - M_*$ scaling relation. Any nonzero σ_μ implies that there is a range of possible masses $\log M_* \pm \sigma_\mu$ for each object of a given black hole mass $M_{\rm BH}$, where we have assumed, for simplicity, a symmetric scatter in the relation. If the number density of galaxies is falling off rapidly in the interval $\log M_* \pm \sigma_\mu$, it will then be more likely to find one of the more numerous small mass galaxies associated with the given black hole and therefore a larger ratio $M_{\rm BH}/M_*$. Thus, given a distribution of galaxy masses and provided that the scatter σ_μ is not too large, the logarithmic offset of each point from the correlation, assumed to be held fixed to the local determination, is given by

$$\Delta \log(M_{\rm BH}/M_*) = 0.67 \times \Delta \log M_{\rm BH} \approx \sigma_\mu^2 \left(\frac{d \log \phi}{d \log M_*} \right)_{\log M_* = (\mu - A)/B} , \qquad (11.14)$$

where $\mu \equiv \log M_{\rm BH}$ and $(A, B) = (1.12, -4.12)$ are slope and intercept of the local scaling relation between BH and host galaxy masses (Häring and Rix 2004).

One can estimate observationally the logarithmic derivative of the galaxy mass function $\frac{d \log \phi}{d \log M_*}$. At $z \approx 2$, the offset expected from such a bias is of the order of 0.25 dex, if $\sigma_\mu = 0.5$ and increases to about 0.5 dex for $\sigma_\mu = 0.7$. The average offset shown in ➋ *Fig. 11-24* is clearly in excess of what is expected in the most extreme case of large intrinsic scatter in the local relation, estimated to be less than 0.5 dex (Gültekin et al. 2009). The data point toward an evolution of the scaling relation, either in normalization or in scatter (or a combination of both).

What are the implications of these findings for our understanding of the cosmological coevolution of black holes and galaxies?

In the next section, we will discuss in more detail a number of physical processes by which AGN can regulate the growth of their host galaxies, thereby affecting any observable evolution of the scaling relations. We will see how, from the physical point of view, a clear distinction has to be made between two modes of AGN feedback.

The first one is associated to the numerous, long-lived, LLAGN, with emitted power dominated by the kinetic energy of their jets and outflows. It becomes increasingly important for very massive holes at low redshift (see ➋ Sect. 3.1.1 above). Many models of galaxy formation invoke such a feedback mechanism in order not to overproduce very massive galaxies in the largest virialized DM halos at low redshift (Croton et al. 2006).

It is not clear, however, how such a feedback mode can effectively couple the SMBH mass with the structural properties of their galactic hosts and give rise to the observed scaling relations. For such a task, modelers have instead turned to feedback modes associated to the phases of fast SMBH growth in bright QSO.

In all feedback models in which the black hole energy injection is very fast (explosive), if strong QSO feedback is responsible for rapidly terminating star formation throughout the entire bulge (Di Matteo et al. 2005), QSOs and, in general, type 1 AGN are associated with the final stage of bulge formation. Then, very little evolution, as well as very little scatter, is expected for the scaling relations, and it is very hard to produce any positive offset like the one observed.

The physics of such ("quasar") mode of AGN feedback remain elusive, as it remains the issue of whether the energy release by the associated process of rapid black hole growth is indeed responsible for halting the conversion of gas into stars on galactic (kpc) scales or whether it is only responsible for a milder form of "self-regulation" by cutting off its own gas supply on nuclear (pc) scales (Hopkins et al. 2009).

4 Cosmology II: AGN Feedback

The phenomenological investigation presented in ➲ Sect. 3 above leaves open the fundamental question about the physical origin of such a clear, parallel differential growth of both the black holes and the galaxy population.

From the discussion of AGN activity in ➲ Chap. 7, it should be immediately apparent that black hole growth is often, if not always, accompanied by the release of enormous amounts of energy, in the form of radiation, outflows, and gravitational waves.

Black holes accreting at high rates in the so-called radiative (or quasar) mode will release of order 10% of the accreted rest mass energy as radiation. They can also drive broad (uncollimated) outflows, again described in more detail in ➲ Chap. 7, and in about 10% of bright AGN, radio emitting, relativistic jets are observed (the quasar 3C273 seems to be such an object that is accreting efficiently *and* making powerful jets at the same time).

But even black holes in the so-called inefficient accretion regime, where cooling is dominated by advective processes rather than radiation, can drive powerful, collimated outflows in the form of relativistic jets. Perhaps the best example of such a powerful "low-efficiency" black hole is the radio galaxy Virgo A, which is the product of inefficient accretion onto the supermassive black hole at the center of M87, which itself is located in the center of the Virgo cluster. See ➲ Chap. 7 for a detailed discussion of the properties of the M87 jet.

This energy will be released directly into the environment from which the black hole grows: the cooling, possibly star-forming gas in the central galaxy. Any transfer of energy to the gas should thus reduce the rate at which gas cools and forms stars. While the direct link between star formation through cooling in the centers of galaxies and black hole growth through accretion is not fully established, it is easy to imagine how such an energy deposition can reduce the rate of accretion onto the central black hole as well.

This process of cooling-induced black hole activity can therefore be considered as a negative feedback loop, in that *increased* accretion activity acts to *decrease* the large-scale gas supply to the black hole. The impact on star formation might be coincidental (if black hole growth is unrelated to the actual star formation rate) or fundamental (if black hole growth is mediated or directly fueled by star formation, e.g., through direct accretion of stars).

Furthermore, while the direct link between black hole growth and star formation suggested by the M–σ_* relation is hidden, and evidence for the suggested underlying feedback process *on stars* is largely circumstantial (as will be discussed below), some important clues can be derived by tracing the evolution of the feedback energy released by growing black holes as a function of black hole mass and redshift.

In this section, we describe in some detail how such an inventory can be made and how feedback itself operates. We will focus primarily on the information about the properties of black holes that can be extracted from observations of feedback, with other chapters discussing the role of black hole growth on the formation of structure in more detail (see ➲ Chap. 6).

4.1 Evidence and Arguments for Feedback

The process of accretion is a multi-scale phenomenon: The range from the place of capture, where the gas first enters the sphere of influence of the black hole, to the event horizon of the black hole spans roughly seven orders of magnitude in scale–too much to simulate in one big simulation for even just one dynamical time on the outer scale.

Yet, as extreme as this range in scales may be, the process of feedback can cover another 5 orders of magnitude more in scale: from the scales of the horizon (about AU size for a typical central black hole in a typical, L_*, galaxy) to scales of entire galaxy clusters (several hundred kiloparsec), a dynamic range of 12 orders of magnitude.

Given that our understanding of accretion is still developing, and that our understanding of jet formation is, at best, elementary, it should not come as much of a surprise that our understanding of AGN feedback is mostly limited to fairly crude statements about energy input and global heating efficiencies from a theoretical perspective.

The best *observational* evidence for feedback is not on galaxy scales at all but on the largest spatial scales on which we can expect black holes to have any meaningful influence: in the centers of galaxy clusters. The reason for this is twofold:

- First, the angular scales on which feedback in galaxy clusters unfolds are readily resolvable by telescopes in all bands of the electromagnetic spectrum.
- Second, the signatures of AGN feedback in galaxy clusters are easy to identify from X-ray and radio imaging, as we will discuss momentarily.

Consequently, we have developed a fairly mature picture of how AGN feedback works on the very large scales and have even successfully simulated the feedback processes in computers.

On smaller scales, the evidence for feedback becomes increasingly circumstantial. Thus, while the link between black hole and galaxy properties may be the most fundamental expression of direct coupling of their growth processes, it is also the most elusive in terms of direct evidence for this coupling.

A number of reasons conspire to limit our observational insight into galactic scale feedback:

- The angular scales of this process are inherently small, given that the feedback must be happening in the centers of galaxies.
- While cluster evolution is happening in the current epoch, at low redshift, galaxy growth happens at higher redshift–during the star formation epoch. This is especially true for the galaxies that seem to require AGN feedback the most.
- Both star formation and rapid black hole growth tend to cloak themselves in dust extinction and photo-electric absorption. It may be that the smoking guns of feedback are mostly hidden behind Compton-thick X-ray absorbers and many magnitudes of dust extinction.

Finally, the tight connection between black hole growth and star formation suggested by the $M-\sigma_*$ relation and by the similarity in the redshift evolution of both populations seems to imply that stars and black holes grow roughly *simultaneously*. For feedback to have a strong impact on star formation and at the same time couple the mass of growing black holes to the mass of stars in a galaxy, one would expect rapid black hole growth to be concurrent with episodes of star formation.

This would imply that feedback on star formation would have occurred during the quasar phases, and since most quasars are radio quiet, this suggests that at least part of the feedback on star formation operates through a different channel than the readily observable “radio galaxy” feedback on cluster scales at low redshift. Given that this feedback must have occurred during the quasar epoch, it is commonly referred to as "quasar mode" feedback.

In fact, the most convincing "evidence" for such a mode comes not from actual observations of black holes but from semi-analytic models of galaxy formation; in order to explain the galaxy luminosity function and galaxy color distribution, modelers have to assume *two* types of feedback: one that disperses and heats the star-forming gas at the end of a star formation cycle (generally triggered by mergers), effectively halting star formation (this is the "quasar mode") and one that maintains the gas in typical elliptical galaxies in its tenuous, hot state (this is the "radio" or "maintenance mode") (Springel et al. 2005; Croton et al. 2006).

However, these models say nothing about the actual physical mechanism of feedback: They assume quasi-spherical heating of the gas in both "quasar" and "radio" mode, and the only thing that distinguishes them is the prescription of how the black hole accretes (whether from cold or hot gas). The more appropriate naming convention is thus "cold" and "hot" mode accretion.

Thus, the circumstantial evidence for "cold" mode feedback does not answer the question of whether the heating/dispersal occurs as a result of winds, jets, or radiation released by the accreting black hole. Given that slowly growing black holes are radiatively inefficient and universally seem to be radio-active (Ho 2008), it has generally been assumed that any feedback from black holes in the "hot" mode must be in the form of jets.

4.2 Feedback in Galaxy Clusters

It is instructive to begin by discussing the obvious examples of AGN feedback. This will inform our discussion of the possible influence of AGN on the process of star formation on galactic scales. In particular, from a discussion of radio galaxy feedback on cluster scales, it is possible to draw quantitative conclusions about "radio-mode" or "hot-mode" feedback by jets from slowly growing black holes. For a more detailed discussion of feedback in galaxy clusters, see McNamara and Nulsen (2007).

On a basic level, the importance of feedback was already apparent with the discovery of powerful radio galaxies in the 1960s and onward (though the relevance of the black hole in this context took longer to establish): radio galaxies, like the example of Cygnus A shown in ➲ *Fig. 11-25*, exhibit diffuse "lobes" of synchrotron emission (see ➲ Chap. 7), on scales of tens and even hundreds of kiloparsec. In other words, the action of a jet from the central black hole deposits magnetized relativistic plasma into the surrounding medium.

Simply summing up the entire synchrotron radiation and making reasonable assumptions about the shape and volume filling fraction of the emitting regions, it is straight forward to derive lower limits on the total energy needed to explain the radio emission. In some cases, the minimum energy derived could be enormous: Perley et al. (1984) found that Cygnus A required at least $E_{\rm min} \gtrsim 10^{60}$ ergs pumped into the lobes by the central black hole. For perspective, this is of the same order as the gravitational binding energy of the Milky Way.

If jets could release this much energy in relativistic gas into the environments of black holes, it would be hard to imagine how the environment could *not* be strongly affected.

The first *direct* evidence for feedback on the gas surrounding the black hole came with the arrival of high-resolution X-ray imaging: Using *ROSAT* data, Boehringer et al. (1993) discovered that the radio galaxy Perseus A (powered by the supermassive black hole in the central cluster galaxy NGC 1275) excavates large cavities in the hot, X-ray emitting thermal gas that fills

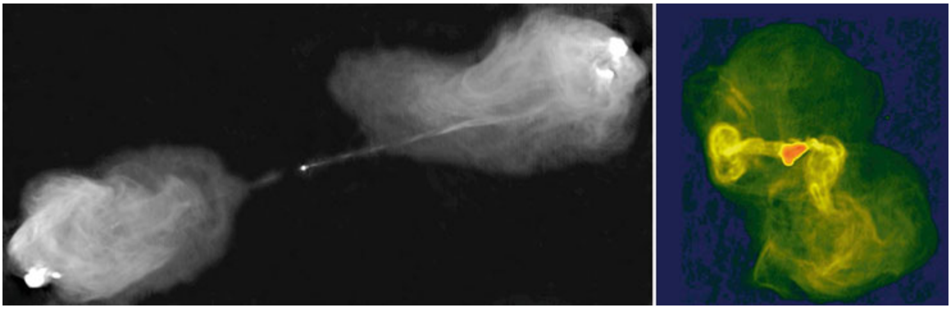


Fig. 11-25
***Left*: The FRII radio galaxy Cygnus A, observed by the VLA at 6-cm; image scale: 150 × 85 kpc; *right*: The large-scale structure of the FR I radio galaxy Virgo A, observed by the VLA at 90 cm. The relativistic inner jet of M87 is contained in the overexposed central radio lobes; image scale: 80 × 80 kpc**

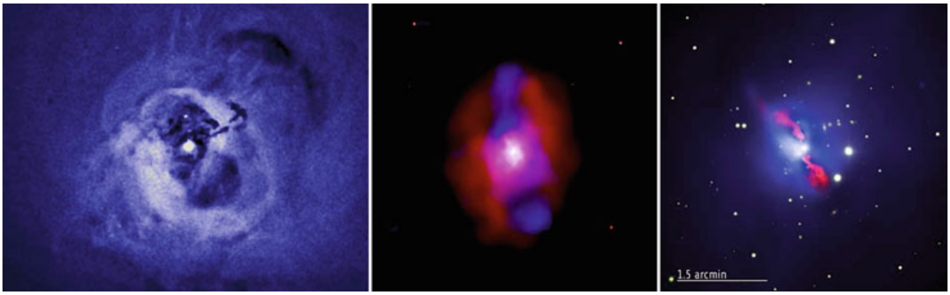


Fig. 11-26
***Left*: Deep *Chandra* observation of the Perseus cluster (Fabian et al. 2006); *middle*: *Chandra* image of MS0735 (*red*) and VLA (*blue*), adopted from McNamara et al. (2005); *right*: *Chandra* image of Hydra A (*blue*) and 6 cm VLA radio image (*red*) (Adopted from NASA (Kirkpatrick et al. 2009); image scale: 80 × 80 kpc)**

the Perseus cluster. The gas is pushed aside into dense shells, and the excavated X-ray cavities are filled with radio emission by the lobes of the radio galaxy (see *Fig. 11-26*).

Similarly, the privileged view we enjoy of the nearby radio galaxy Virgo A (also known as M87) allowed a uniquely detailed study of its multi-scale emission well before the idea of feedback had taken hold. The inner (roughly kpc) jet of M87 is discussed in some detail in Chap. 7. However, lower frequency observations revealed a much richer picture on scales just outside of the visible galaxy, still in the very center of the Virgo cluster (Owen et al. 2000): curling and twisting strands of radio emission, connecting the nucleus to a set of radio lobes about 20 kpc in radius, and misaligned with respect to the central jets by about 90° on the sky.[6]

[6] Part of this misalignment could be due to projection, of course, given that at least the inner jet is directed fairly close to the line of sight.

Both Perseus and Virgo are cool core clusters, and in particular, Perseus had long been considered a prototypical example of a cooling flow. That is, the radiative cooling time in the center of the cluster is shorter than the age of the cluster. In a quasi-hydrostatic model of a cluster (inward gravity, mostly provided by the dark matter contribution, balanced by an outward thermal pressure gradient), this would imply that the cluster must be contracting on a Kelvin–Helmholtz timescale, with gas at the center cooling rapidly to star-forming temperatures. Cooling gas from further out in the cluster would replace the gas in a slow, subsonic inflow (for a review, see Fabian 1994, and references therein).

Even before the era of *Chandra* and *XMM*, it was already apparent that this simple picture of ongoing inflow of cooling gas did not accurately describe cool core clusters: The implied rates for star formation were an order of magnitude higher than the observed rates.

Radio surveys of cluster centers revealed that essentially all traditional "cooling flow" clusters had active radio galaxies in their centers (Burns 1990). Generally, these radio sources are Fanaroff–Riley type I sources (henceforth abbreviated as FR I; see ➋ Chap. 7 for a discussion of radio source morphology), though it is not entirely clear whether this is due to lower average source power compared to field FR II galaxies or due to the increased gas density in clusters (frustrating source evolution and possibly leading to increased entrainment).

Guided by the detailed examples of feedback in the Virgo and Perseus clusters and the observed mismatch between the X-ray cooling rate and the star formation rates in clusters, the first models of black hole feedback in the context of galaxy clusters were presented in Tabor and Binney (1993).

4.2.1 The *Chandra* and *XMM-Newton* View

The role of AGN in regulating the cooling of gas in cool core clusters was brought into clear focus with the launches of *Chandra* and *XMM-Newton* in two ways:

Chandra observations revealed the presence of cavities just like those found in the center of Perseus in virtually every cool core cluster, providing the observational confirmation that the radio galaxies present in these clusters *actively perturb* the gas. Generically, the cavities appear to be surrounded by relatively cool gas.[7] Deep *Chandra* observations sometimes reveal multiple cavities on different scales, which has been interpreted as evidence for variability in the AGN power.[8]

At the same time, *XMM-Newton's* high-resolution X-ray spectra of the cluster centers revealed that radiative cooling must be impeded below a threshold temperature of order 10 million degrees (about 1 keV), at a temperature where cooling should be efficient and rapid due to the flattening of the cooling curve for thermal gas as atomic line cooling becomes dominant (Peterson et al. 2003).

This result is consistent with the observed lack of star formation in central cluster galaxies, compared to the hundreds of solar masses of star formation per year that would have been expected based on simple cooling flow models. It moves the discrepancy of cool gas missing at *molecular* temperatures to cool gas missing at *X-ray* temperatures below about a keV. In either case, a heating agent is needed, but in the "revised" cooling flow problem, the gas that must be

[7] This was surprising because one might naively expect the gas most strongly affected by feedback to be hot.

[8] But see Morsony et al. (2010) for arguments why the presence of cavities is not a sufficient argument for AGN duty cycles.

preferentially targeted is at about a keV and will thus occupy a much larger volume fraction, which should make it easier to interact with for any feedback mechanism.

Thus, after about a decade of study, a standard paradigm has emerged from the high incidence of radio loud AGN in cool core clusters and from the theoretical requirement of a heating agent that maintains the temperature distributions in galaxy clusters: radio galaxies provide the energy needed to counterbalance cooling in the centers of clusters.

From these observations, and from theoretical modeling (Begelman and Cioffi 1989; Reynolds et al. 2001), a simple understanding of radio source evolution has been developed that underpins the radio galaxy feedback paradigm. According to this picture, radio source evolution separates into three stages (e.g., Reynolds et al. 2001):

1. In the initial supersonic phase, jet plasma inflates cocoons that are strongly overpressured relative to the environment. These cocoons must expand, and the rapid energy release implies that this expansion is supersonic in the frame of the environment.

   The expansion is similar to that of a wind-blown bubble described by Castor et al. (1975): The cavity radius roughly follows the self-similar scaling[9]

   $$R \propto \left(\frac{P_{\mathrm{jet}}\,t^{3}}{\rho_{\mathrm{ICM}}}\right)^{1/5} \tag{11.15}$$

   which is functionally equivalent to the Sedov–Taylor solution if the blast energy is replaced by the injected energy over time, $P_{\mathrm{jet}}\,t$. In this expression, ρ_{ICM} is the density of the environment.

   Note that this argument neglects the actual jet propagation completely and assumes the jet energy is randomized as the jet encounters the environment (in more powerful FR II sources) or through entrainment (which has been suggested as the dissipation agent in FR I source; Laing and Bridle 2002).
2. As the source expands, the pressure inside the cocoon and in the shell eventually approaches the pressure of the environment and the expansion becomes subsonic. The initially generated shock wave of the supersonic expansion will continue to coast outward, leaving behind a subsonically expanding cavity. As the expansion velocity becomes subsonic, the confinement is dominated by the thermal pressure of the environment, so the solution changes to a pressure confined bubble,

   $$R \propto \left(\frac{P_{\mathrm{jet}}\,t}{p_{\mathrm{e}}}\right)^{1/3} \tag{11.16}$$
3. Finally, as the source expansion velocity drops below the buoyancy speed or once the pressure of the source drops below the dynamic pressure of motions in the environment, the cavities/cocoon will detach and float away from the black hole buoyantly or advectively.

   Once cold gas has refilled the central region of the cluster, any jet activity will start the cycle anew.

It is clear from (➋ 11.15) that jet power and density are the controlling variables in the initial evolution: lower power jets in denser environments will be more easily frustrated (with more slowly expanding cocoons that become sub-sonic and unstable at smaller sizes).

[9] This expression applies to bubbles smaller than a cluster pressure scale height. It is straight forward to extend it to stratified power-law atmospheres.

4.2.2 Estimating the Kinetic Power of a Radio Source

Before the discovery of X-ray cavities, measurements of the *kinetic* power of radio galaxies (i.e., the total power traveling down the jets) were limited to estimates based on the observed synchrotron emission (see discussion in ❯ Sect. 4.2). These were hampered by several factors:

- Without knowledge of the field strength, an observed synchrotron flux, along with an estimate of the emitting volume could only provide a lower limit of the total energy in the radio plasma (essentially assuming equipartition between the energy in electrons and magnetic field).
- The estimate of the volume depends strongly on the volume filling fraction, which is not measurable.
- Synchrotron aging can cause electrons to cool and develop a sharp cutoff in the synchrotron spectrum. Thus, significant amounts of energy in lower energy electrons would be unaccounted for in the total power budget.

One of the most important results from the discovery of X-ray cavities in clusters is a robust, independent way to estimate the power of cluster radio sources. It is based on the fluid mechanics of inflated bubbles: In order to inflate a cavity in the intra-cluster gas, the jet must (a) displace the material in the environment into a shell surrounding the cavity, which is of the order of $E_{pV} \sim pV$ (and depends on the details of the inflation history of the bubble) and (b) replace it with relativistic, magnetized gas. At a minimum, the amount of energy needed to do this is the work done on the cluster gas and the internal energy of the radio plasma. If the expansion of the cavity is adiabatic, the total energy needed is

$$E_{\text{cavity}} = \frac{\gamma p V}{(\gamma - 1)} \sim 4pV, \tag{11.17}$$

where γ is the adiabatic index of the gas inside the cavity and is typically assumed to be $\gamma = 4/3$, given the presence of relativistic electrons and tangled magnetic fields.

Estimating the age of the cavity, and thus the jet power needed to inflate it, is more difficult and introduces some uncertainty. Direct kinematic measurements of the expansion velocities of cavities are impossible with current X-ray telescopes. However, in many cases, the observations suggest that the temperature of the shells surrounding the cavities is *low*. In this case, it is safe to assume that the recent expansion of the cavity was subsonic. Thus, the sound crossing time of the bubble radius is a reasonable *lower* limit on the cavity age:

$$t_{\text{cavity}} \geq \tau_{\text{sonic}} = \frac{R_{\text{cavity}}}{c_{\text{s}}}, \tag{11.18}$$

where R_{cavity} is the radius of the cavity and c_{s} the sound speed of the cluster gas.

Given that most cavities are found in cluster *centers*, a reasonable *upper* limit on the age of the cavity is the buoyant rise time τ_{buoy} of the bubble in the gravitational potential of the cluster (since any bubble older than τ_{buoy} would have risen out of the cluster center):

$$t_{\text{cavity}} \leq \tau_{\text{buoy}} \sim \frac{2R_{\text{cavity}}}{v_{\text{buoy}}} \sim \frac{2R_{\text{cavity}}}{c_{\text{s}}\sqrt{\frac{4}{3}\frac{d\ln(P)}{d\ln R}\frac{1}{C_{\text{W}}}}} = 2\tau_{\text{sonic}}\sqrt{\frac{3C_{\text{W}}}{4}\frac{d\ln(R)}{d\ln(P)}} \sim 2\tau_{\text{sonic}}, \tag{11.19}$$

where C_{W} is the drag coefficient of the rising cavity and typically assumed to be of order $C_{\text{W}} \sim 0.5$ and the exact numerical value of the expression under the square root depends on the pressure scale height of the cluster in bubble radii but should be of order unity for typical observed bubble radii.[10] For *detached* cavities, the appropriate age to use is the buoyancy time for the projected distance instead of the bubble diameter.

Given that buoyancy sets in after the source becomes subsonic, this loosely brackets the power inferred for the jets from the measurement of cluster cavities to

$$\frac{E_{\text{cavity}}}{\tau_{\text{buoy}} + \tau_{\text{sonic}}} \sim \frac{E_{\text{cavity}}}{3\tau_{\text{sonic}}} \lesssim P_{\text{jet}} \lesssim \frac{E_{\text{cavity}}}{\tau_{\text{sonic}}}. \tag{11.20}$$

Estimates of the central cluster density and temperature (and thus pressure and sound speed) are readily obtained from X-ray images and spectra. The most difficult part is the estimate of the cavity volume, since errors in the estimated cavity radius and viewing angle uncertainties can compound to errors of up to an order of magnitude in source power. Nonetheless, this method has afforded us with a large number of reliable estimates of jet powers to within a factor of a few for dozens of radio galaxies in nearby clusters.[11]

While early, shallower *Chandra* exposures of cluster centers only showed cool shells in the vicinity of the cavities, deep observations of a number of important clusters later also showed the presence of shocks surrounding at least some of the cavities (e.g., McNamara et al. 2005; Wise et al. 2007; Forman et al. 2007). Given the generic picture of how radio sources evolve over time described in ➤ Sect. 4.2.1, the presence of *weak* shocks should be expected (and had been predicted in Reynolds et al. 2001).

In fact, an important corollary from (➤ 11.15) is that the initial strongly supersonic phase is short lived in typical cluster environments: The expansion velocity of the shell is

$$v_{\text{shell}} = \frac{dR}{dt} \propto \left(\frac{P_{\text{jet}}}{\rho_{\text{ICM}} R^2}\right)^{1/3} \propto R^{-2/3} \propto t^{-2/5} \tag{11.21}$$

Given that cavities should remain in the cluster centers for about 2–3 sound crossing times before becoming buoyant, the fraction $f_{>M}$ of time a given radio galaxy spends expanding supersonically at or above a given Mach number M, relative to the total dynamic lifetime before buoyant removal, should only be of order

$$f_{>M} \sim \frac{\tau_{>M}}{3\tau_{\text{sonic}}} \sim \frac{1}{3} M^{-5/2}. \tag{11.22}$$

Thus, the observational lack of evidence for sources expanding at large Mach numbers does not rule out that radio sources go through this strong shock phase. It is, however, short lived and only a small mass fraction of AGN's environment passes through a strong shock.

The detection of a shock (which requires not just the detection of a surface brightness jump but also a temperature jump, which is most easily identified through a harder X-ray color) offers a significantly better diagnostic of jet power than the cavity method: Because the shock strength is an indication of the expansion velocity, a measured shock radius, brightness, and strength can be modeled using simple 1D spherical shock models to give a reliable source power. Typical Mach numbers for shocks detected in clusters are between one and two, consistent with this argument.

[10] The buoyancy speed can never exceed the sound speed.

[11] It should be kept in mind that the inferred powers are averages over the cavity age, which can be between millions to hundreds of millions of years old.

This rather simple parametric description of radio source evolution is complemented by a growing body of numerical simulations of jet-driven feedback. Initial 2-dimensional simulations generally supported the simple picture (Reynolds et al. 2001). However, a thorough understanding requires full 3-dimensional simulations, and work on understanding the details of feedback ab initio is still in the early stages.

A key problem posed by 2D and early 3D simulations is the apparent contradiction of highly bipolar release of energy in the jet and the need for mostly isotropic heating (Vernaleo and Reynolds 2006). While outgoing shocks assume a spheroidal shape relatively quickly, heating by shocks is insufficient to solve the cooling flow problem (see (➲ 11.22)). In addition, jets evacuate cocoons around them. Ongoing, unidirectional jet activity was found to propagate inside this cocoon, dynamically and energetically isolated from the inner cluster gas and thus unable to counteract cooling in the center.

The solution might lie in the interaction with the cluster and in the internal dynamics of jets: Heinz et al. (2006) found that the jets can continue to efficiently couple to the inner cluster if their axes are subjected to a moderate wobble and, crucially, the cluster itself is dynamically evolved in the context of a cosmological simulation. Simulations generally show that the complex X-ray appearance of clusters, and even the appearance of multiple successive cavities, can be generates by a single, ongoing episode of jet activity and a dynamic cluster atmosphere (Morsony et al. 2010). This suggests that the interaction between jets and clusters goes both ways: Cluster dynamics affects jet dynamics and vice versa.

4.2.3 Cluster Radio Sources as a Population

A census of nearby clusters with cavities reveals a number of important insights about the statistical and global properties of *central* cluster radio sources:

- When plotting jet power vs. X-ray luminosity, radio sources straddle the heating=cooling line: About half of the radio sources sampled have kinetic powers that are higher than the cooling luminosity of the cluster (see ➲ *Fig. 11-27* adapted from Rafferty et al. 2006).
- More importantly, the jet power appears to be correlated with the cooling luminosity of the cluster: More powerful radio sources are found in clusters with higher cooling rates. This is exactly the signature one would expect in an AGN feedback picture (Rafferty et al. 2006).
- The jet power measured from cavities appears to be related to the Bondi accretion rate of the cluster. That is, in clusters with higher central densities and/or lower core temperatures, the jet power is higher (Allen et al. 2006). While this correlation does not imply that black holes actually accrete at the Bondi rate in clusters, the relatively close match suggests that black holes accrete directly from the cluster gas as it flows into the central galaxy (see ➲ *Fig. 11-27*).

Together, these results suggest that the balance between heating and cooling in clusters is relatively tight. In fact, the high efficiencies suggested by these findings, in a global sense, as well as from studies of individual, powerful black holes with relatively low masses, have led to the suggestion that the black holes powering the jets in cluster centers might even require the extraction of black hole spin (McNamara et al. 2011).

However, given the significant uncertainties in the jet power estimates, and given that clusters are, in all likelihood, not steady state systems (thus, cooling rates can temporarily exceed heating rates, as long as they are balanced on average), a sufficient region of parameter space

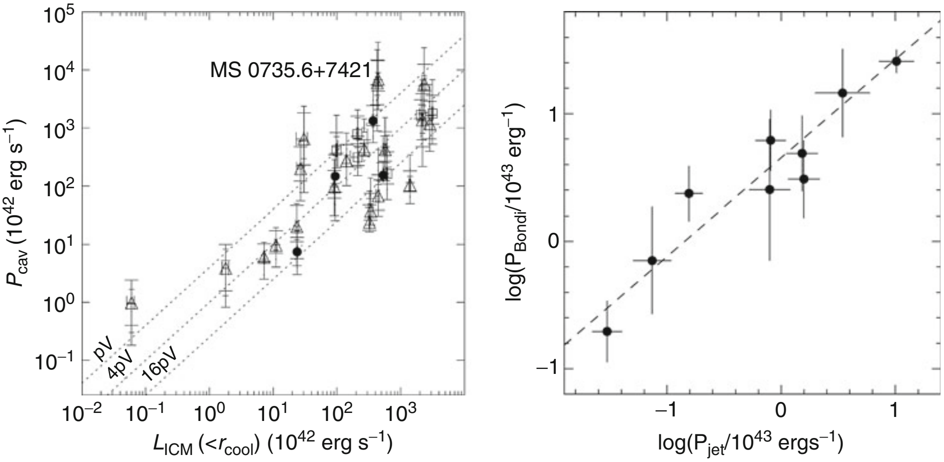


Fig. 11-27

***Left*: jet power inferred from cluster cavities plotted against X-ray cooling luminosity from within the cluster cooling radius, adopted from Rafferty et al. (2006); the *diagonal lines* indicate the efficiency required for the jet power to offset cooling, with the line labeled as 4*pV* equivalent to 100% efficiency, adopted from Rafferty et al. (2006); *right*: jet power plotted against Bondi accretion rate (Adopted from Allen et al. (2006))**

is still allowed in which AGN can balance cooling without requiring extreme efficiencies, accretion rates, or black hole masses.

All of the detailed studies of radio galaxy feedback in clusters have been limited to clusters at relatively low redshifts, because of the need for high-fidelity X-ray images of cavities and shocks. In addition, high-redshift cluster samples are sparse. Consequently, constraints on cluster feedback at higher redshift are much harder to obtain. Radio surveys of high-redshift clusters do indicate an increase in the cluster radio luminosity function (Branchesi et al. 2006), hinting at an increase in feedback activity.

Finally, it is worth mentioning that, while the central massive elliptical galaxies harbor by far the most massive black holes in clusters, other cluster galaxies can be radio loud. This situation typically leads to the formation of a bent radio source. The exploration of the statistics of non-central cluster radio sources is ongoing, and it has been suggested that this population of sources could contribute to cluster feedback (Hart et al. 2009). However, it is difficult to envision a scenario whereby a population of such sources will dominate the heating rate in the average cluster. It is also unclear how they could respond to central gas cooling as would be required for thermal regulation of the cluster gas.

4.3 Feedback in Groups

Due to the lower temperatures and densities in the intragroup medium, detecting evidence for group-wide feedback from X-ray observations is significantly more difficult than in clusters.

Statistical X-ray studies of cluster and group samples show that, generically, lower mass halos have excess central entropy when compared to self-similar models of halo formation, indicating that an additional source of nongravitational heating must have injected entropy preferentially into the low mass systems (e.g., Ponman et al. 2003). AGN have been suggested as a possible heating mechanism (e.g., Short and Thomas 2009). ➲ *Figure 11-28* shows how the inclusion of feedback in detailed models of cluster and group atmospheres affects primarily low mass systems and raises their central entropies.

And, in fact, surveys of groups suggest that such a scenario might work: Dong et al. (2010) find that a fraction of at least 25% of the groups in their survey contain clearly detectable cavities, with a clear preference for cavities to be found in groups with *cooler* cores (as is the case in clusters). In their sample, the presence of cavities does not appear to be correlated with the 1.4-GHz radio flux.

On the other hand, a survey of radio properties of groups shows that the central temperatures of groups with central radio sources are *elevated* compared to a radio quiet sample of groups, which was suggested as a possible indication of ongoing radio galaxy feedback (Croston et al. 2005).

4.4 Radio-Mode Feedback in Galaxies

The best direct evidence for black holes affecting the surrounding gas within galaxies comes again from combined radio and *Chandra* observations of nearby objects: ➲ *Fig. 11-29* shows images of three nearby galaxies where jets clearly excavate cavities and affect gas on sub-galactic scales: M87, M84, and NGC5128 (Centaurus A). In the latter case, the entrainment of gas into the jet, and the formation of a strong shock driven by the expanding southwestern radio lobe is directly visible in X-rays.

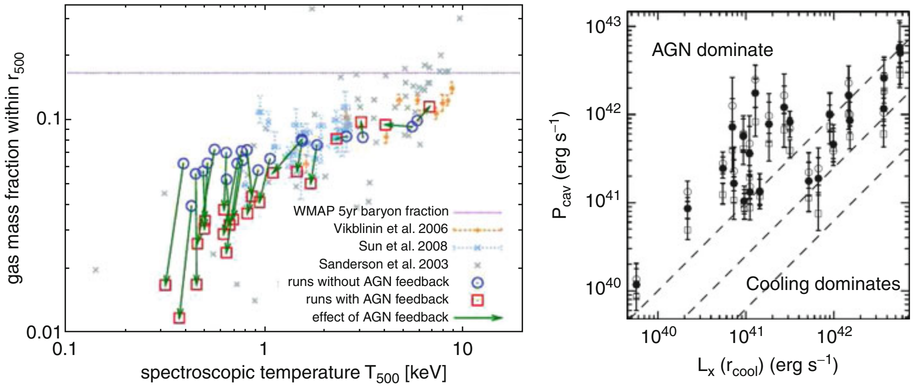


Fig. 11-28

***Left*: Simulations show the increasingly important effect of AGN feedback on lower mass clusters and groups; plotted is the central gas mass fraction as a function of virial temperature (mass); adopted from Puchwein et al. (2008); *right*: same plot as in the *left panel* of ➲ *Fig. 11-27* but for atmospheres of elliptical galaxies instead of clusters (Adopted from Nulsen et al. (2007))**

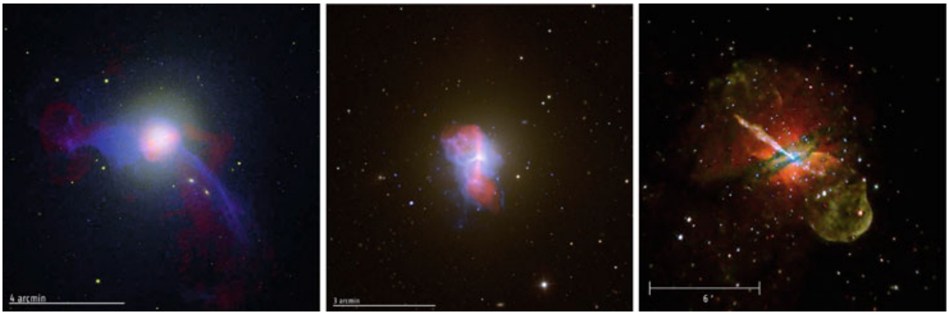


Fig. 11-29
***Left*: multiwavelength image of the central region of the Virgo cluster, showing the influence of the black hole on sub-galactic scales; *middle*: multiwavelength image of the galaxy M84 in the Virgo cluster (*blue*: radio, *red*: X-ray, *yellow*: visible); *right*: multi-wavelength image of Centaurus A**

These are clear local examples of relatively powerful, evolved radio galaxies. Numerous other examples have been studied individually. Yet, perhaps the most convincing argument for the *importance* and prevalence of jets in massive galaxies come from statistical studies.

In a study of X-ray cavities in and around nearby ellipticals, Nulsen et al. (2007) found that AGN power more than matched the cooling luminosity of the gas in the hot galactic X-ray halo; compared to the same study in clusters (shown in the left hand panel of *Fig. 11-27*), the galaxies fall consistently above the heating=cooling line (McNamara and Nulsen 2007).

The implication that radio-mode feedback is important in massive galaxies (those that have detectable X-ray halos) is complemented by statistical studies of the radio source incidence in galaxies of different type and mass: The left panel of *Fig. 11-30* shows a steady increase of the fraction of radio loud galaxies (i.e., galaxies above a fixed radio luminosity per stellar mass) with stellar mass of the host galaxy (Best et al. 2005). The most massive galaxies exhibit the highest radio fluxes *per stellar mass*, indicating that radio-mode feedback is most active in today's most massive galaxies, with 10% or more of the most massive galaxies hosting radio loud AGN.

Given that more massive galaxies harbor more massive black holes, one might naively expect this result to suggest that radio loudness is a fraction of black hole mass. Early studies seemed to suggest this (e.g., Franceschini et al. 1998). However, deep radio surveys of a wider class of black holes show that the radio loudness, defined relative to the bolometric flux, actually *increases* for *decreasing* Eddington ratios (Ho 2002): The right panel of *Fig. 11-30* shows that radio emission increases in relative brightness (compared to the bolometric luminosity) for lower luminosity AGN.

This suggests that black holes become relatively *more* efficient at liberating energy in the form of jets as their luminosity (and presumably accretion rate) drops, such that *all* black holes at sufficiently low accretion rate appear to be driving some form of radio loud outflow.

The idea that low-luminosity black holes are universally radio loud arose roughly in parallel also in the study of X-ray binary black holes (Gallo et al. 2003), where it is possible to track individual black holes across outburst and decline into quiescence. These observations showed that a jet was always present at low luminosities, with increasing *relative* radio flux at decreasing X-ray luminosities.

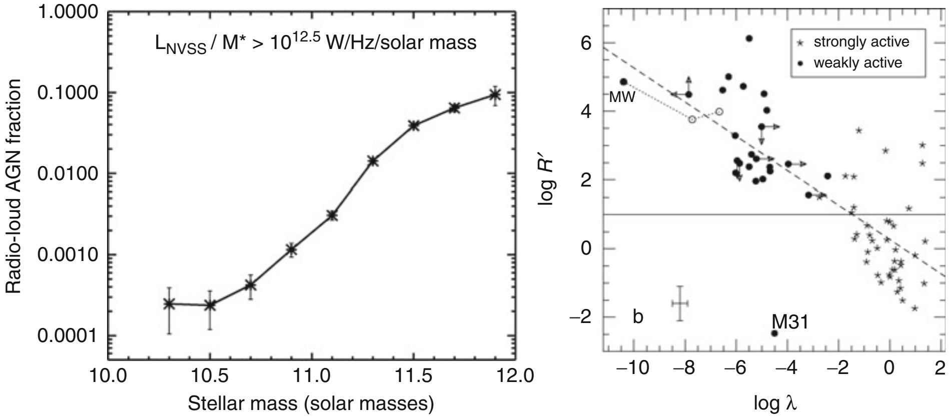


Fig. 11-30

***Left*: Fraction of radio loud objects (defined as objects with a radio luminosity larger than $10^{12.5}$ W Hz^{-1} per solar mass of stars of its host galaxy) plotted against stellar mass of the host, adopted from Best et al. (2005); *right*: Radio loudness *R*′ (here defined as 5-GHz radio flux to 2,500 Åflux) as a function of Eddington ratio *λ* (defined as the ratio of bolometric AGN flux to Eddington flux) (Adopted from Ho (2002))**

While a thorough theoretical understanding is still missing, it has been shown that low-efficiency accretion leads to the formation of geometrically thick flows, as opposed to the geometrically thin accretion discs found in, for example, Seyfert galaxies and quasars. In such a geometrically thick (quasi-spherical) flow, it might be much easier to build up significant magnetic flux, even just from stochastic turbulent dynamo processes in the disc, which could in turn drive the jet (see Chap. 7).

Anecdotally, the case of the M87 jet makes the perfect illustration of this point: The total radiative output from the black hole is unimpressive[12] at $L_{\mathrm{M87}} \sim 10^{41}$ ergs s^{-1}, about two orders of magnitude below the luminosity corresponding to efficient accretion at the Bondi rate of that particular black hole. Meanwhile, the estimated jet power from this object is about two orders of magnitude larger (Forman et al. 2007), consistent with the accretion power one would derive if the black hole were accreting at the Bondi accretion rate, thus making M87 a jet-dominated low-luminosity black hole.

While a direct translation of radio loudness into feedback efficiency is challenging, this result suggests that black hole feedback might be more prevalent than just in the most powerful AGN in the most massive galaxies. In fact, the ubiquity of low-luminosity AGN and their increased dominance at low redshift (see *Fig. 11-8*) supports both the idea of a radio maintenance mode of ongoing feedback in early-type galaxies (Croton et al. 2006) as well as the picture of gentle feedback (through buoyancy rather than shocks) not just in clusters but also in galaxies.

Additionally, ongoing low-level jet activity seems much better suited at targeting gas within the galaxy than explosive outbursts in the form of powerful radio galaxies that would transport most of the energy far beyond galactic scales, and the prominent examples of radio-mode

[12] Much of this radiation may actually be in the form of X-rays from the unresolved base of the jets itself.

feedback shown in *Fig. 11-29* should be considered extreme outliers from more typical radio-mode feedback. This statement can, in fact, be quantified based on the global AGN evolutionary models outlined in Sect. 3.1.1, which we shall discuss next.

4.5 Quantifying the Efficiency of the Radio Mode

The observed omnipresence of radio cores[13] in low-luminosity AGN and the observed increase in radio loudness of X-ray binaries at low luminosities can be placed on a solid theoretical footing. Jets launch in the innermost regions of accretion flows around black holes, and at low luminosities, these flows likely become mechanically (i.e., advectively) cooled.

Such flows can, to lowest order, be assumed to be scale invariant: A low-luminosity accretion flow around a ten solar mass black hole, accreting at a fixed, small fraction of the Eddington accretion rate, will be a simple, scaled-down version of the same flow around a billion solar mass black hole (with the spatial and temporal scales shrunk by the mass ratio). It follows, then, that jet formation in such a flow should be similarly scale invariant.

This assumption is sufficient to derive a very generic relation between the radio luminosity emitted by such a scale invariant jet and the total (kinetic and electromagnetic) power carried down the jet, *independent* of the unknown details of how jets are launched and collimated (Heinz and Sunyaev 2003): The synchrotron radio luminosity L_ν of a self-absorbed jet core depends on the jet power $P_{\rm jet}$ through

$$L_{\rm radio} \propto P_{\rm jet}^{\frac{17+8\alpha}{12}} M^{-\alpha} \sim P^{\frac{17}{12}}. \tag{11.23}$$

where M is the mass of the black hole and $\alpha \sim 0$ is the observable, typically flat radio spectral index of the synchrotron power law emitted by the core of the jet. This relation is a result of the fact that the synchrotron photosphere (the location where the jet core radiates most of its energy) moves further out as the size scale and the pressure and field strength inside the jet increase (corresponding to an increase in jet power). As the size of the photosphere increases, so does the emission. The details of the power-law relationship are an expression of the properties of synchrotron emission.

For a given black hole, the jet power should depend on the accretion rate as $P_{\rm jet} \propto \dot{M}$ (this assumption is implicit in the assumed scale invariance). On the other hand, the emission from optically thin low-luminosity accretion flows itself depends nonlinearly on the accretion rate, roughly as $L_{\rm acc} \propto \dot{M}^2$, since two body processes like bremsstrahlung and inverse Compton scattering dominate, which depend on the square of the density. Thus, at low accretion rates, $L_{\rm radio} \sim L_{\rm bol}^{\frac{17}{24}}$, which implies that black holes should become more radio loud at lower luminosities (Heinz and Sunyaev 2003; Merloni et al. 2003; Falcke et al. 2004). It also implies that more massive black holes should be relatively more radio loud than less massive ones, at the same *relative* accretion rate $\dot{M}/M$.

Equation 11.23 is a relation between the observable core radio flux and the underlying jet power. Once calibrated using a sample of radio sources with known jet powers, it can be used to estimate the jet power of other sources based on their radio properties alone (with appropriate provisions to account, statistically, for differences in Doppler boosting between different sources).

[13] The "core" of a jet is the brightest innermost region of the jet, where the jet just becomes optically thin to synchrotron self absorption, that is, the synchrotron photosphere of the jet.

The cluster radio sources shown in ➡ *Fig. 11-27* provide such a sample. Plotting the core (unresolved) radio power against the jet power inferred from cavity and shock analysis shows a clear nonlinear relation between the two variables (Merloni and Heinz 2007). Fitting this relation provides the required constant of proportionality and is consistent (within the uncertainties) with the power-law slope of 17/12 predicted by (➡ 11.23):

$$P_{\mathrm{jet}} = P_0 \left(\frac{L_{\mathrm{radio}}}{L_0}\right)^{\zeta} \sim 1.6 \times 10^{36}\,\mathrm{ergs\,s^{-1}} \left(\frac{L_{\mathrm{radio}}}{10^{30}\,\mathrm{ergs\,s^{-1}}}\right)^{0.81} \tag{11.24}$$

with an uncertainty in the slope ζ of 0.11, where $L_{\mathrm{radio}} = \nu L_\nu$ is measured at $\nu = 5\,\mathrm{GHz}$.

Because this relation was derived for the cores of jets, which display the characteristic flat self-absorbed synchrotron spectrum, care has to be taken when applying it to a sample of objects: Only the core emission should be taken into account, while extended emission should be excluded. As discussed in ➡ Sect. 2.1.1, radio luminosity functions are separated spectrally into flat and steep sources, and we can use both samples to limit the contribution of flat-spectrum sources from both ends.

Given a radio luminosity function Φ_{rad} and an appropriate correction for relativistic boosting, (➡ 11.24) can be used to derive the kinetic luminosity function of jets (Heinz et al. 2007; Merloni and Heinz 2008):

$$\Phi_{\mathrm{kin}}(P_{\mathrm{jet}}) = \Phi_{\mathrm{rad}} \left[L_0 \left(\frac{P_{\mathrm{jet}}}{P_0}\right)^{\frac{1}{\zeta}}\right] \frac{1}{\zeta} \frac{L_0}{P_0} \left(\frac{P_{\mathrm{jet}}}{P_0}\right)^{\frac{1-\zeta}{\zeta}}. \tag{11.25}$$

The resulting kinetic luminosity functions for the flat-spectrum radio luminosity functions[14] from Dunlop and Peacock (1990) and de Zotti et al. (2005) are plotted in the right panel of ➡ *Fig. 11-31*.

Since the figure plots $P \cdot \Phi_{\mathrm{P}}$, the curves show *directly* the total contribution of AGN at a given jet power to the total feedback power at a given redshift. At the low-luminosity end, these curves are roughly flat, implying that low-luminosity source contributed a significant fraction of the total power. These are the low-luminosity AGN presumably responsible for radio-mode feedback, and they dominate the total jet power output at low redshift.

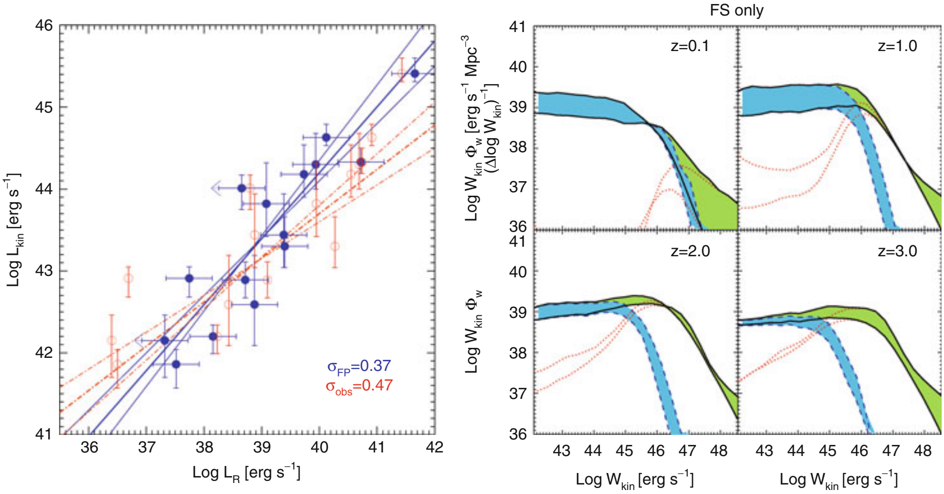
Integrating the luminosity function over P_{jet} gives the local jet power density ρ_{Pjet}, which, at redshift zero, is of the order of $\langle \rho_{\mathrm{Pjet}} \rangle \sim 6 \times 10^{39}\,\mathrm{ergs\,s^{-1}\,Mpc^{-3}}$, which is comparable to the local power density from supernovae, but will be significantly above the supernova power in early-type galaxies (which harbor massive black holes prone to accrete in the radio mode but no young stars and thus no type 2 supernovae).

Finally, integrating Φ_{kin} over redshift gives the total kinetic energy density u_{Pjet} released by jets over the history of the universe, $u_{\mathrm{Pjet}} \sim 3 \times 10^{57}\,\mathrm{ergs\,Mpc^{-3}}$. By comparing this to the local black hole mass density ρ_{BH}, we can derive the average conversion efficiency η_{jet} of accreted black hole mass to jet power:

$$\eta \equiv \frac{u_{\mathrm{Pjet}}}{\rho_{\mathrm{BH}} c^2} \approx 0.2\% - 0.5\%. \tag{11.26}$$

In other words, about half a percent of the accreted black hole rest mass energy gets converted to jets, *averaged* over the growth history of the black hole.

[14] Comparison to the steep-spectrum luminosity function shows that the error in Φ_P from the sources missed under the steep-spectrum luminosity function is at most a factor of 2.



■ Fig. 11-31
***Left*: *Red open circles* show jet power (measured from X-ray cavities) plotted against 5-GHz core radio luminosity (also shown in *blue solid circles* is a Doppler boosting corrected version of the same points) along the power-law fit given in (❯ 11.24); adopted from Merloni and Heinz (2007); *right*: Kinetic jet luminosity function for different redshift bins (*green curve* shows all radio sources, *blue* shows radio sources in the radio mode, *red* shows radio sources in quasar mode) (Adopted from Merloni and Heinz (2008))**

Since most black hole mass was accreted during the quasar epoch, when black holes were mostly radio quiet, about 90% of the mass of a given black hole was accreted at zero efficiency (assuming that only 10% of quasars are radio loud). Thus, the average jet production efficiency during radio loud accretion must be at least a factor of 10 higher, about 2–5%, comparable to the *radiative* efficiency of quasars. These are exactly the kinds of efficiencies needed for radio-mode feedback to work.

4.6 Quasar Mode Feedback

Arguments for black hole feedback on galactic scales stem primarily from three facts:

- The deviation of the bright end of the galaxy luminosity function from self-similar predictions (Springel et al. 2005), that is, a dearth of bright galaxies.
- The bimodality of the galaxy distribution in color-magnitude space (Strateva et al. 2001), with early-type galaxies forming the "red sequence" and late type galaxies forming the "blue cloud." Between these two populations lies, naturally, the so-called "green valley," which is relatively sparsely populated. This bimodality is shown in ❯ *Fig. 11-32*.
- The tight relation between stellar bulge mass and black hole mass, which suggests a common formation scenario. Given that massive black holes grew mostly as quasars during the epoch of star formation, this suggests a relationship between both. In fact, the argument for quasar mode feedback was first made in part to motivate black hole–galaxy scaling relations (e.g., Silk and Rees 1998; Wyithe and Loeb 2003).

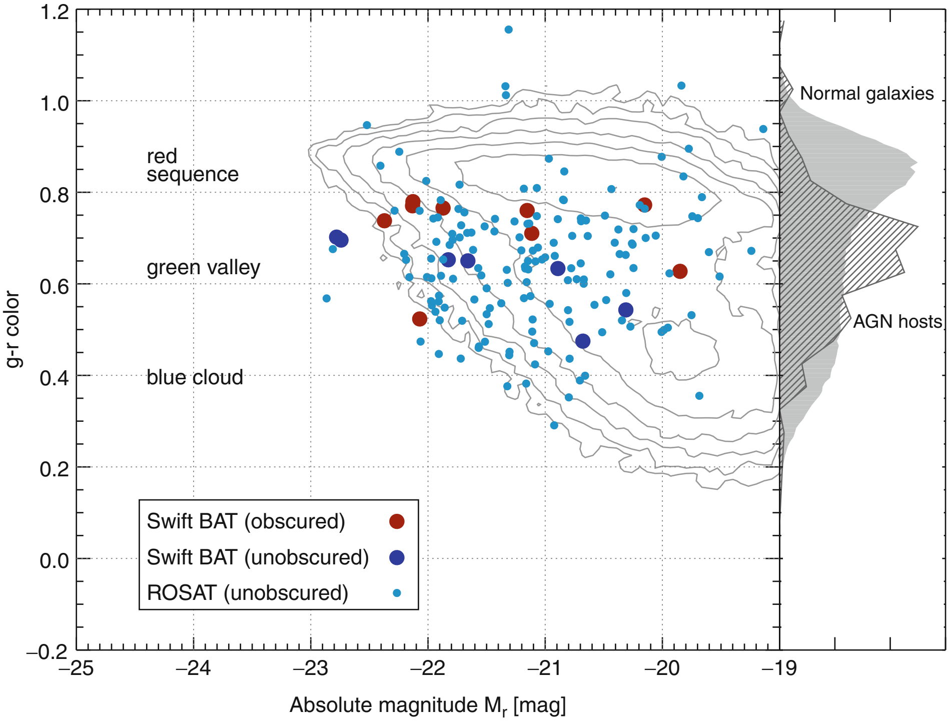


Fig. 11-32

Contours of galaxy counts as a function of R-band magnitude and g–r color, showing the existence of two separate populations – the red sequence on *top* and the blue cloud on the *bottom*. Overplotted are the locations of X-ray selected AGN (hard and soft), with a clear preference for a location in the "green valley" in between both populations (Adopted from Schawinski et al. (2009))

Rapidly growing black holes are attractive as agents of feedback on ongoing star formation because they have similar growth histories, they can be found in the centers of galactic bulges (the stellar populations their feedback is supposed to influence), they can release large amounts of energy isotropically, and they are likely to be fueled rapidly in response to galaxy mergers, which also trigger star formation.

Numerical simulations of black hole feedback in individual galaxy mergers have produced impressive visualizations of how rapid, isotropic energy injection by a growing black hole can heat and disperse the cool, star-forming gas, in essence explosively terminating star formation and black hole growth (Di Matteo et al. 2005). In part as a result of these successes, quasar mode feedback is now routinely incorporated into cosmological simulations of structure formation and semi-analytic models of galaxy formation (Croton et al. 2006; Bower et al. 2006).

In these simulations, the prescription of how black holes accrete is simplified to variations of the Bondi accretion rate, necessitated by the unresolvably vast dynamic range of the problem. Energy is injected isotropically, which is an appropriate zero-order choice given our lack of knowledge about the actual channel through which the energy is delivered. What the simulations tell us is that efficient black hole feedback *can* regulate star formation and black hole growth. But because black hole feedback and supernova feedback are operationally very similar and because the presumed AGN feedback mechanism is generic, it is difficult to extract more detailed information about the AGNs themselves from the models.

In addition, the causality of the interaction of black holes with the star-forming gas surrounding them is not yet fully established. It is also plausible that star formation itself provides the feedback through supernovae and that competitive accretion starves both black holes and stars, leading to a passive link between black holes and stars.

Identifying currently ongoing episodes of feedback has proven to be difficult, in part because of the large degrees of visible and soft X-ray extinction toward star-forming regions and because of the small angular scales involved. Proving the *causal* connection between AGN activity and terminated star formation is even more difficult.

Generally, one would assume that galaxies caught in the act of feedback should just start their transition from the blue cloud to the red sequence, as the population of recently formed early stars fades without any replenishment. The relative under-density of galaxies in the green valley suggests that this transition is a relatively rapid process (one would expect it to occur roughly on A-star life times).

Stellar population modeling has successfully been used to identify such post-starburst galaxies. And indeed, sources have been found among this class that show clear evidence for very fast outflows in excess of $1,000\,\mathrm{km\,s^{-1}}$ (Tremonti et al. 2007) that might be the smoking gun. Estimating the mass in the outflow has proven to be difficult, and we have to await deep imaging that can directly resolve the outflow to quantify the energetic impact of the AGN on the galactic gas.

In addition, surveys of (hard) X-ray selected AGN find these sources to preferentially lie in the green valley (from the all-sky *Swift-BAT* survey; Schawinski et al. 2009), as can be seen in ➋ *Fig. 11-32*. Since AGN accretion timescales can be expected to be shorter than the transition time across the green valley, this observation suggests that the AGN activity comes *after* star formation has been terminated. Since this is true also for hard X-ray selected AGN, this conclusion should not be affected by obscuration unless AGN in the act of feedback are so heavily obscured that even *Swift* cannot detect them.

Thus, the question of what is at the heart of the putative quasar mode feedback is left unanswered. Generally, AGN can release energy via three channels: through radiation, through jets, and through uncollimated outflows (i.e., "winds").

The most obvious source of feedback energy in efficiently accreting black holes is, of course, the radiation itself. Since most bright AGN are obscured, we can infer that a high fraction of the initially emitted light is reabsorbed by the surrounding gas. If some of the energy is deposited on sufficiently large scales (rather than into gas bound to the black hole), it can in principle supply the energy for feedback. Models of radiative feedback (Ciotti and Ostriker 2001; Sazonov et al. 2005) generally rely on Compton heating. The efficiency of radiative feedback requires about 10% of the radiation to be absorbed on scales outside of the Bondi radius but within the star-forming region of the host. Whether radiative transfer will always conspire to provide such an arrangement is an open question.

Efficiently accreting AGN (quasars and Seyferts) are also known to generate massive winds: Optical absorption line studies show outflows at velocities of thousands of kilometers per second. The most dramatic demonstration comes from the class of broad absorption line quasars (BAL QSOs), which show high column densities of absorption in visible and X-rays (indicating high mass fluxes) and wind velocities up to $50{,}000\,\mathrm{km\,s^{-1}}$.

If column density measurements made in the X-ray band trace the same gas that produces the large outflow velocities in optical absorption lines, the power and mass contained in these winds would be of the same order as the total radiative power of these objects (Furlanetto and Loeb 2001). Driving such a wind would presumably require some form of mechanical input (e.g., from magneto-centrifugal launching as described in Chap. 7) in addition to radiative driving.

The ubiquity of outflow signatures in efficiently accreting AGN, coupled with the large wind efficiencies inferred from the more extreme cases, has made AGN winds the primary mechanism invoked in feedback models (see, e.g., King 2005). Given the uncertainties in column density of the high-velocity gas, a direct imaging detection of outflow signatures (like the cavities in the case of AGN feedback in clusters) would provide more certainty that winds can affect the surrounding gas on the scales needed for feedback to operate.

The uncertainty about whether winds are powerful enough to drive feedback raises the interesting question whether episodes of *powerful jet* activity in quasars can lead to feedback on galactic scales and whether they can be observed. In the simple framework of radio source dynamics laid out in Sect. 4.2.1, an episode of jet activity will inflate a supersonically expanding cocoon, the size scale of which depends on the jet power and the density of the environment.

In the dense environments of star-forming regions, one might thus expect sources to go through a compressed evolution, with slowed or even stalled expansion as sources run into dense gas. In such a scenario, the initial expansion might produce strong shocks (given the cold gas they encounter) but at much reduced shock temperatures given the slower expansion. Is it possible that powerful radio sources in dense environments can heat the gas sufficiently to provide the quasar mode feedback postulated by semi-analytic models?

Given that about 10% of all powerful quasars are radio loud and given that the required *average* feedback efficiency for the quasar mode can be an order of magnitude lower still, jet powered feedback may actually contribute significantly to the quasar mode as well. In fact, a class of sources that might represent these powerful radio quasars in the act of feedback exists in the so-called compact-steep-spectrum (CSS) and the gigahertz-peaked-spectrum (GPS) sources (O'Dea 1998).

These are small-scale radio sources that show clear signs of strong absorption to the radio spectrum (indicating a high local pressure and thus ISM density) and otherwise appear similar to classical radio sources but on smaller scales. The cause for their compactness has been debated since their discovery: They might be young sources, in the very early stages of supersonic expansion, or frustrated older sources, caught in very dense environments. In either case, this would be a population of sources directly heating the dense gas in the centers of galaxies, where quasar mode feedback is observed.

Recent evidence does suggest that these sources are indeed young and that we are looking at infant powerful radio sources (e.g., Holt et al. 2008; Kunert-Bajraszewska et al. 2010). The high rate of incidence, compared to bona fide quasars, suggests that they are a short-lived phenomenon, which would make them effective short-cycle thermostats in a feedback scenario.

The detection of compact radio sources in high-redshift star-forming environments seems to support the role of jets in quasar mode feedback: A number of high-power compact radio sources have been found in actively star-forming regions with powerful outflows (Nesvadba et al. 2007, 2011) and in dense, high-z cooling flow environments (Siemiginowska et al. 2010). Because we know jet feedback works in the context of clusters and likely in the "maintenance" mode of feedback, and because we know that CSS and GPS sources are (a) frequent and (b) powerful, they present an attractive alternative to the wind-driven QSO mode of feedback.

Simulations of jets in dense, multiphase environments, as might be expected in star-forming galaxies, already show significant promise in solving the question of how bipolar, highly collimated jets in even very powerful radio sources could efficiently heat the gas in galaxies (Wagner and Bicknell 2011).

5 Cosmogony

We have seen in ➋ Sect. 3.1 that the total mass density estimated in relic supermassive black holes at $z \sim 0$ is consistent with the total mass accreted by growing black holes during (obscured and unobscured) AGN phases for a radiative efficiency of the accretion process ($0.06 < \epsilon_{\rm rad} < 0.2$, depending on the bolometric corrections and local mass density exact estimate), well in line with the prediction of classical relativistic accretion disc models (Novikov and Thorne 1973).

In fact, the validation of the Soltan (1982) argument implies that the last few e-folds of a SMBH's mass are mainly grown via (radiatively efficient) classical accretion discs, rather than through mergers or radiatively inefficient accretion. If this is true, however, the very process of cosmological black hole growth through accretion quickly erases the initial condition, namely, the primordial mass function of seed black holes, making it almost impossible to deduce the physical properties of early black hole formation from observations probing redshifts smaller than that corresponding to the most efficient growth ($z \approx 2-3$). That is, unless a specific range of BH masses is identified which is less affected by the complex process of AGN activation during structure formation.

Indeed, some have argued that small mass black holes in isolated, small mass galaxies could have maintained a "memory" of the seeding mechanism (in their location with respect to the scaling relations defined for more massive systems, for example), being less affected by the multiple generations of hierarchical mergers in the emerging cosmic web (see, e.g., Volonteri 2010). Very few observational constraints are available for this class of objects, however.

On the other hand, the observation of luminous quasars at $z \simeq 6$ (e.g., Fan et al. 2001) has shown that it is possible to probe directly the earlier epochs of massive black hole assembly and thus to try to directly constrain the various physical processes responsible for planting the seeds that grow into the giant monsters in the nuclei of galaxies.

5.1 The First Black Holes: Observational Constraints and Theoretical Ideas

The observed luminosity functions of AGN suggest a rapid decline of the total luminosity density above $z \approx 3$ (see ➋ *Fig. 11-18*). At face value, the constraints on the very-high-redshift evolution of the population at $z > 5$ come primarily from bright optical quasars detected in very large area surveys, but recent indications from large area and moderately deep radio (Wall et al. 2005) and X-ray (Civano et al. 2011) surveys do provide a consistent picture for the evolution of the most luminous AGN over all observational wavebands (➋ *Fig. 11-33*).

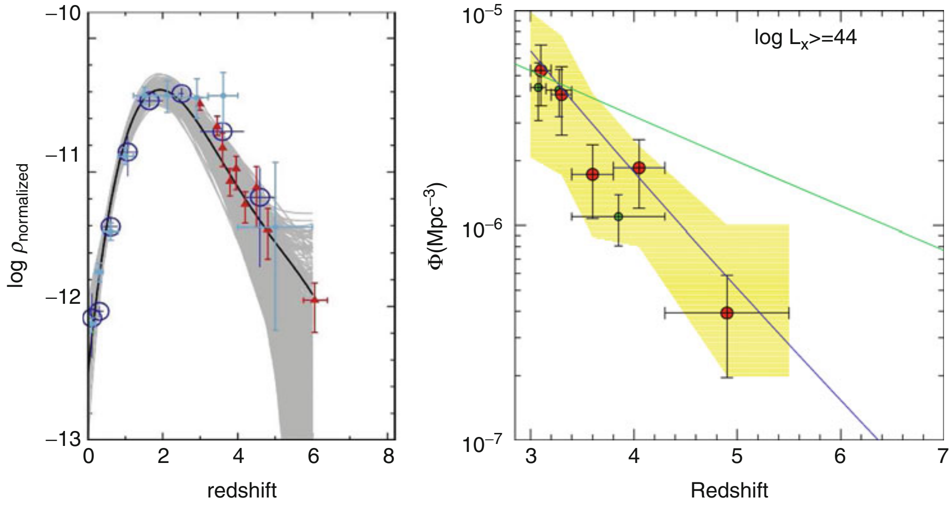


Fig. 11-33

Observational constraints on the decline of QSO number densities at high redshift. *Left*: Relative space density of QSOs (ρ) as a function of redshift. The *shaded area* and *black line* represent the current QSO space-density determination from the Parkes quarter-Jansky flat-spectrum sample of radio-selected QSOs. *Light-blue-filled* and *dark-blue open circles* show the soft X-ray data from ROSAT, Chandra, and XMM-Newton surveys. Space-density behavior of optically selected QSOs is given by the set of *dark-red triangles*. The X-ray and optical QSO data were scaled vertically to match the current determination of space density at redshifts 2–2.5. From Wall et al. (2005). *Right*: The comoving space density at bright 2–10-keV X-ray luminosity from the Chandra-COSMOS survey, computed taking into account the effect of obscuration. The *blue curve* corresponds to the X-ray selected AGN space density computed for the same luminosity limit from models of the CXRB. The *yellow-shaded* area represents the maximum and minimum space density. The *green symbols* correspond to the data of XMM-COSMOS (From Civano et al. (2011))

The observed rapid decline of QSO number density toward high redshift translates into a rapid decrease in the number of AGN-generated ionizing photons toward the end of the re-ionization epoch. Accurate determination of the QSO rest-frame UV luminosity function are thus crucial to assess the role of growing black hole might have played in re-ionizing the universe. Current estimates suggest that galaxies do dominate the comoving emissivity of ionizing photons esaping in the intergalactic medium at $z > 4$ (Haardt and Madau 2012, and references therein).

Despite their rarity, very-high-redshift QSOs can provide interesting constraints on the early evolution (and even formation mechanisms) of nuclear supermassive black holes. The high metal enrichment observed in high-z AGNs (see Sect. 1.1), even in those close to reionization ($z \sim 6$), indicates that the host galaxies of these AGNs must have undergone a powerful and rapid burst of star formation. And indeed, vigorous star formation is observed in such high-z quasars, as inferred from the detection of prominent PAH features, strong far-IR emission and from the detection of the [CII]158-m line (see Maiolino 2009), redshifted to millimeter wavelengths. Most likely, we are witnessing the coeval, rapid formation of massive bulges along with their supermassive central black holes.

These luminous quasars detected at $z > 6$, when the universe was less than 1 Gyr old have estimated BH masses (from the "virial method") in excess of $\approx 10^9 M_\odot$, and it is by no means a trivial task to grow such massive holes in the relatively short time available. Assuming continuous growth at an Eddington ratio of $\lambda = L/L_{\rm Edd}$:

$$\frac{dM_{\rm BH}}{dt} = (1-\epsilon_{\rm rad})\lambda \times L_{\rm Edd}/(\epsilon_{\rm rad}c^2), \tag{11.27}$$

we have, for the final BH mass as a function of the initial mass,

$$M_{\rm BH,f}(t) = M_i \exp\left[(1-\epsilon_{\rm rad})/\epsilon_{\rm rad} \times (t/\tau_{\rm salp})\right], \tag{11.28}$$

where we have defined the typical e-folding time (the so-called Salpeter time) as

$$\tau_{\rm salp} = \frac{\lambda c \sigma_{\rm T}}{4\pi G m_{\rm p}} = 0.45\left(\frac{\lambda}{1}\right)\,\text{Gyr}. \tag{11.29}$$

Depending on the redshift of formation of the seed of mass M_i, and on the average radiative efficiency of the accretion process, only a limited range of final BH masses can be reached at $z = 6$, as shown in *Fig. 11-34* (Shapiro 2005). If the BH seed masses are in the range expected from Pop III remnants, of the order of a few hundred solar masses, then highly radiative efficient

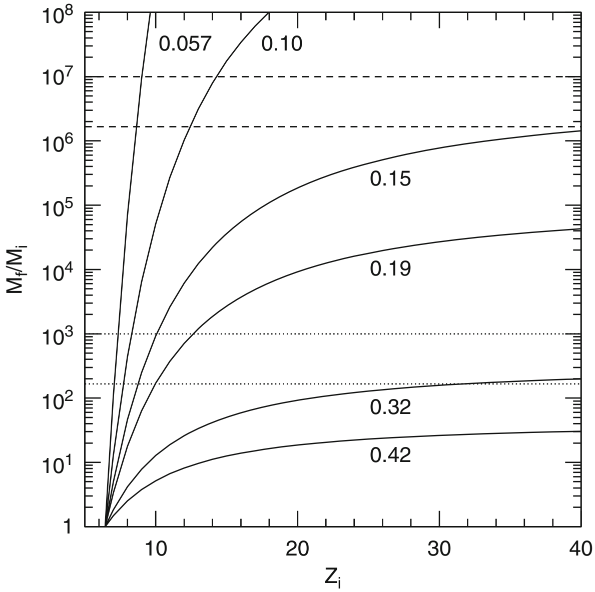


Fig. 11-34
Ratio of final-to-initial black hole mass at $z = 6$ calculated from (11.28) assuming $\lambda = 1$ (i.e., continuous Eddington-limited accretion) as a function of the formation redshift of the seed black hole. Each *solid curve* is labeled with the corresponding value of the radiative efficiency $\epsilon_{\rm rad}$. *Dotted horizontal lines* show the ratio implied by the observed $z \sim 6$ QSOs for massive seeds, while the dashed ones that for stellar mass (Pop III; see Sect. 5.1 for details) (From Shapiro (2005))

($\epsilon_{\text{rad}} > 0.1$) accretion is excluded, as it would not allow enough mass to be accumulated into the black hole rapidly enough.

Such a scenario has a number of major difficulties. First of all, if the accretion is indeed continuous and proceeds at high Eddington rates, the accretion flow should form a geometrically thin, optically thick disc. If this is the case, even a black hole with zero initial dimensionless angular momentum $a_i = 0$ will be rapidly spun up by the angular momentum captured with the accreting gas. Bardeen (1970) has shown that in this case, the final spin obeys:

$$a_f = \frac{r_{\text{ms,i}}^{1/2}}{3} \frac{M_i}{M_f} \left[4 - \sqrt{3 r_{\text{ms,i}} \left(\frac{M_i}{M_f} \right)^2 - 2} \right]. \tag{11.30}$$

where $r_{\text{ms,i}}$ is the radius of the marginally stable orbit of the initial black hole (itself a monotonic function of spin, with $r_{\text{ms}}(a = 0) = 6$ and $r_{\text{ms}}(a = 1) = 1$ in gravitational units). Thus, an initially non-spinning BH is spun up by accretion of gas in a classical geometrically thin and optically thick accretion disc as soon as[15] $M_f/M_i = \sqrt{6}$. Since an accretion disc around a maximally rotating hole will radiate with an efficiency $\epsilon_{\text{rad}} = 0.42$, we are left with the impossibility of growing black holes larger than a few thousands of solar masses via prolonged coherent accretion onto a stellar mass seed (see ➔ *Fig. 11-34*).

A few solutions have been proposed to the above problem:

1. First, accretion might not be coherent, but rather stochastic, such that the accreting gas comes into the gravitational sphere of influence of the black hole in parcels with randomly oriented angular momenta. If the mass of each parcel is small enough (roughly speaking, if $\Delta M \ll \sqrt{6} M_i$), then the BH spin vector performs a random walk, but as it is easier to spin a black hole down than up, the net effect is a relatively low average spin (and correspondingly lower radiative efficiency) of the final hole (King and Pringle 2006).
2. In the primordial cosmological setup where protogalaxies form, the first black holes grow in a very gas-rich environment, and there is no reason to believe that the gas could not flow toward the central black holes at vastly super-Eddington rates. It is not clear, however, whether the hole can swallow matter that fast. On the one hand, quasi-spherical inflows can be established, where the radial velocity of the accreting gas is so high that the photons produced inside the disc by the viscous torques cannot escape (Frank et al. 2002). In those cases, although the emerging luminosity is at most logarithmically in excess of the Eddington limit, the accretion rate onto the hole can be orders of magnitude larger. On the other hand, the accretion flow can start blowing out matter at the (large) radius where the locally produced energy exceeds the local Eddington limit (Shakura and Sunyaev 1973). A powerful wind ensues, which may prevent the mass accretion rate onto the black hole from exceeding the Eddington limit by more than a factor of a few.
3. Black hole seeds in the early universe can be more massive than the remnants of Pop III stars. These would have formed by direct collapse of primordial massive stars (see, e.g., Volonteri 2010). For example, some theoretical models have argued that the gas chemistry in the most massive, hottest primordial DM halos can prevent fragmentation of the cooling gas and lead to the formation of very massive stars. Their cores will rapidly collapse, leaving a black hole

[15] Note that the above calculation assumes that there is no torque at the inner boundary of the accretion disc (Novikov and Thorne 1973). Magnetic linkage between the disc, the plunging region, and the event horizon can modify the above picture, reducing the maximal spin a BH can reach (Krolik et al. 2005). Nonetheless, most numerical models of geometrically thin magnetized discs are still consistent with a rapid spin of the BH.

at the center of a quasi-static, radiation pressure-supported supermassive star. The resulting object, called a quasistar, resembles a red giant with a luminosity comparable to a Seyfert nucleus. The black hole grows inside it until the cooling photosphere can no longer sustain its own radiation pressure and the envelope disperses, leaving behind the naked seed black hole of typically 10^4–10^5 $M_\odot$ (Begelman 2010).

4. Finally, a large number of major mergers could help relax the demands on the efficiency and stability of gas accretion on the first black holes by enhancing the final-to-initial mass ratio by a factor of the order of the number of equal-mass (or major) mergers along the main tree of the hierarchy.

   Cosmological numerical simulations (Li et al. 2007) provide a possible route to the formation of a 10^9 $M_\odot$ at $z \sim 6$ by starting with Pop III stellar mass seeds at $z \sim 30$ which experience an early phase of continuous, Eddington-limited accretion (subject, however, to the limitations discussed above), before entering the merger tree at $z \sim 14$, where a large number of major merger events is able to accumulate the final mass, even in the presence of AGN feedback.

Analytic and semi-analytic models of the early assembly of massive black holes that include (with varying degrees of sophistication) the many competing processes (mergers, gravitational wave recoil and nuclear black hole ejection, dynamical friction on wandering BHs, pristine gas accretion, etc.) make clear predictions for the early seed mass distributions, the initial conditions for SMBH growth that we would like to probe. ➲ *Figure 11-35* shows one such prediction, comparing the outcomes of three different formation scenarios: direct collapse, runaway stellar mergers in high-redshift stellar clusters, and Population III remnants. A more complete census

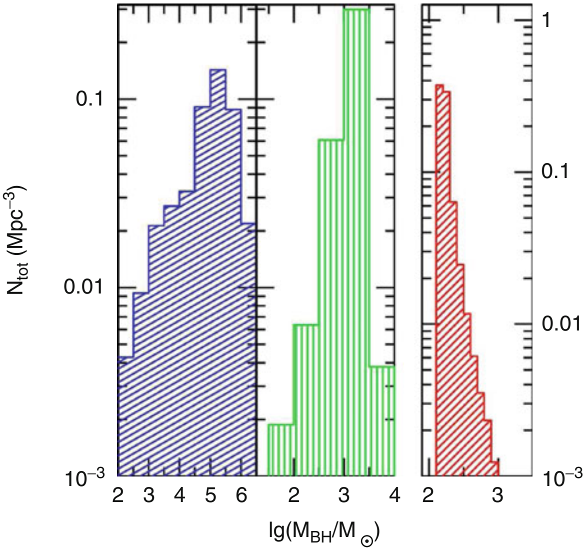


Fig. 11-35

Mass function of seed massive Black Holes for three different formation scenarios: direct collapse (*left*), runaway stellar mergers in high-redshift stellar clusters (*center*) and Population III remnants (*right*). Note the different y-axis scale for the Pop III case (From Volonteri (2010))

of the AGN population at $z \sim 6$ or even a few detections of $z \sim 10$ AGN with the next generation of large astronomical facilities could provide direct means to distinguish among these simple formation scenarios, allowing us to glimpse into the obscured epoch when the first nuclear black holes formed.

Acknowledgments

We would like to thank warmly our Editor William C. Keel for the kind invitation to write this chapter in the new edition of the "Planets, Stars and Stellar System" book, and for the patience and endurance he demonstrated in putting up with our delays. We acknowledge the many contributions and suggestions from a vast number of colleagues and collaborators. In particular, we wish to thank Viola Allevato, Roberto Assef, Silvia Bonoli, Niel Brandt, Marcella Brusa, Johannes Buchner, Nico Cappelluti, Scott Croom, Francesco De Gasperin, Guenther Hasinger, Phil Hopkins, Brian McNamara, Paul Nulsen, Eric Perlman, Chris Reynolds and Gordon Richards. This material is based upon work supported by the National Science Foundation under Grant No. 1109347. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. AM's work was partially supported by the DFG Cluster of Excellence "Origin and Structure of the Universe".

Cross-References

- Active Galactic Nuclei
- Clusters of Galaxies
- Galaxies in the Cosmological Context
- Large Scale Structure of the Universe

References

Allen, S. W., Dunn, R. J. H., Fabian, A. C., Taylor, G. B., & Reynolds, C. S. 2006, MNRAS, 372, 21
Allevato, V., et al. 2011, ApJ, 736, 99
Antonucci, R. 1993, ARA&A, 31, 473
Assef, R. J., et al. 2011, ApJ, 728, 56
Bardeen, J. M. 1970, ApJ, 161, 103
Barger, A. J., et al. 2005, AJ, 129, 578
Begelman, M. C. 2010, MNRAS, 402, 673
Begelman, M. C., & Cioffi, D. F. 1989, ApJL, 345, L21
Bennert, V. N., et al. 2011, ApJ, 742, 107
Best, P. N., Kauffmann, G., Heckman, T. M., Brinchmann, J., Charlot, S., Ivezić, Ž., & White, S. D. M. 2005, MNRAS, 362, 25
Boehringer, H., Voges, W., Fabian, A. C., Edge, A. C., & Neumann, D. M. 1993, MNRAS, 264, L25
Bonoli, S., Marulli, F., Springel, V., White, S. D. M., Branchini, E., & Moscardini, L. 2009, MNRAS, 396, 423
Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., & Lacey, C. G. 2006, MNRAS, 370, 645
Boyle, B. J., & Terlevich, R. J. 1998, MNRAS, 293, L49
Branchesi, M., Gioia, I. M., Fanti, C., Fanti, R., & Perley, R. 2006, A&A, 446, 97
Brandt, W. N., & Hasinger, G. 2005, ARA&A, 43, 827
Brusa, M., et al. 2010, ApJ, 716, 348
Burns, J. O. 1990, AJ, 99, 14
Cappelluti, N., et al. 2009, A&A, 497, 635
Castor, J., McCray, R., & Weaver, R. 1975, ApJL, 200, L107

Chaudhary, P., Brusa, M., Hasinger, G., Merloni, A., & Comastri, A. 2010, A&A, 518, A58+

Ciotti, L., & Ostriker, J. P. 2001, ApJ, 551, 131

Civano, F., et al. 2011, ApJ, 741, 91

Cole, S., et al. 2001, MNRAS, 326, 255

Cowie, L. L., Songaila, A., Hu, E. M., & Cohen, J. G. 1996, AJ, 112, 839

Croom, S. M., et al. 2009, MNRAS, 399, 1755

Croston, J. H., Hardcastle, M. J., & Birkinshaw, M. 2005, MNRAS, 357, 279

Croton, D. J., et al. 2006, MNRAS, 365, 11

Danese, L., Franceschini, A., Toffolatti, L., & de Zotti, G. 1987, ApJL, 318, L15

de Zotti, G., Massardi, M., Negrello, M., & Wall, J. 2010, A&AR, 18, 1

de Zotti, G., Ricci, R., Mesa, D., Silva, L., Mazzotta, P., Toffolatti, L., & González-Nuevo, J. 2005, A&A, 431, 893

Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604

Dong, R., Rasmussen, J., & Mulchaey, J. S. 2010, ApJ, 712, 883

Dunlop, J. S., & Peacock, J. A. 1990, MNRAS, 247, 19

Fabian, A. C. 1994, ARA&A, 32, 277

Fabian, A. C., & Iwasawa, K. 1999, MNRAS, 303, L34

Fabian, A. C., Sanders, J. S., Taylor, G. B., Allen, S. W., Crawford, C. S., Johnstone, R. M., & Iwasawa, K. 2006, MNRAS, 366, 417

Falcke, H., Körding, E., & Markoff, S. 2004, A&A, 414, 895

Fan, X., et al. 2001, AJ, 121, 54

Ferrarese, L., & Merritt, D. 2000, ApJL, 539, L9

Forman, W., et al. 2007, ApJ, 665, 1057

Franceschini, A., Vercellone, S., & Fabian, A. C. 1998, MNRAS, 297, 817

Frank, J., King, A., & Raine, D. 2002, Accretion Power in Astrophysics (3rd ed.; Cambridge, UK: Cambridge University Press). ISBN 0-521-62957-8, 2002, XIV + 384 pp.

Fukugita, M., & Peebles, P. J. E. 2004, ApJ, 616, 643

Furlanetto, S. R., & Loeb, A. 2001, ApJ, 556, 619

Gallo, E., Fender, R. P., & Pooley, G. G. 2003, MNRAS, 344, 60

Gebhardt, K., et al. 2000, ApJL, 539, L13

Giacconi, R., Gursky, H., Paolini, F. R., & Rossi, B. B. 1962, Phys. Rev. Lett., 9, 439

Gilli, R., Comastri, A., & Hasinger, G. 2007, A&A, 463, 79

Gültekin, K., et al. 2009, ApJ, 698, 198

Haardt, F., & Madau, P. 2012, ApJ, 746, 125

Hamann, F., & Ferland, G. 1992, ApJL, 391, L53

Häring, N., & Rix, H. 2004, ApJL, 604, L89

Hart, Q. N., Stocke, J. T., & Hallman, E. J. 2009, ApJ, 705, 854

Hasinger, G. 2008, A&A, 490, 905

Hasinger, G., Miyaji, T., & Schmidt, M. 2005, A&A, 441, 417

Heinz, S., & Sunyaev, R. A. 2003, MNRAS, 343, L59

Heinz, S., Brüggen, M., Young, A., & Levesque, E. 2006, MNRAS, 373, L65

Heinz, S., Merloni, A., & Schwab, J. 2007, ApJL, 658, L9

Ho, L. C. 2002, ApJ, 564, 120

Ho, L. C. 2008, ARA&A, 46, 475

Holt, J., Tadhunter, C. N., & Morganti, R. 2008, MNRAS, 387, 639

Hopkins, P. F., Richards, G. T., & Hernquist, L. 2007, ApJ, 654, 731

Hopkins, P. F., Murray, N., & Thompson, T. A. 2009, MNRAS, 398, 303

Iwasawa, K., & Taniguchi, Y. 1993, ApJL, 413, L15

Juarez, Y., Maiolino, R., Mujica, R., Pedani, M., Marinoni, S., Nagao, T., Marconi, A., & Oliva, E. 2009, A&A, 494, L25

King, A. 2005, ApJL, 635, L121

King, A. R., & Pringle, J. E. 2006, MNRAS, 373, L90

Kirkpatrick, C. C., Gitti, M., Cavagnolo, K. W., McNamara, B. R., David, L. P., Nulsen, P. E. J., & Wise, M. W. 2009, ApJL, 707, L69

Krolik, J. H., Hawley, J. F., & Hirose, S. 2005, ApJ, 622, 1008

Kunert-Bajraszewska, M., Gawroński, M. P., Labiano, A., & Siemiginowska, A. 2010, MNRAS, 408, 2261

Laing, R. A., & Bridle, A. H. 2002, MNRAS, 336, 328

Li, Y., et al. 2007, ApJ, 665, 187

Longair, M. S. 1966, Nature, 211, 949

Longair, M. S. 2008, Galaxy Formation, ed. M. S. Longair (Berlin: Springer)

Madau, P., Ferguson, H. C., Dickinson, M. E., Giavalisco, M., Steidel, C. C., & Fruchter, A. 1996, MNRAS, 283, 1388

Magorrian, J., et al. 1998, AJ, 115, 2285

Maiolino, R. 2009, in Astronomical Society of the Pacific Conference Series, Vol. 408, The Starburst-AGN Connection, eds. W. Wang, Z. Yang, Z. Luo, & Z. Chen (San Francisco: Astronomical Society of the Pacific), 235–+

Marconi, A., Risaliti, G., Gilli, R., Hunt, L. K., Maiolino, R., & Salvati, M. 2004, MNRAS, 351, 169

Martini, P., & Weinberg, D. H. 2001, ApJ, 547, 12

Massardi, M., Bonaldi, A., Negrello, M., Ricciardi, S., Raccanelli, A., & de Zotti, G. 2010, MNRAS, 404, 532

Mateos, S., et al. 2008, A&A, 492, 51

McNamara, B. R., & Nulsen, P. E. J. 2007, ARA&A, 45, 117

McNamara, B. R., Nulsen, P. E. J., Wise, M. W., Rafferty, D. A., Carilli, C., Sarazin, C. L., & Blanton, E. L. 2005, Nature, 433, 45

McNamara, B. R., Rohanizadegan, M., & Nulsen, P. E. J. 2011, ApJ, 727, 39

Merloni, A., & Heinz, S. 2007, MNRAS, 381, 589

Merloni, A., & Heinz, S. 2008, MNRAS, 388, 1011

Merloni, A., Heinz, S., & di Matteo, T. 2003, MNRAS, 345, 1057

Merloni, A., et al. 2010, ApJ, 708, 137

Morsony, B. J., Heinz, S., Brüggen, M., & Ruszkowski, M. 2010, MNRAS, 407, 1277

Nesvadba, N. P. H., Lehnert, M. D., De Breuck, C., Gilbert, A., & van Breugel, W. 2007, A&A, 475, 145

Nesvadba, N., et al. 2011, MNRAS, 415, 235

Novikov, I. D., & Thorne, K. S. 1973, in Black Holes (Les Astres Occlus), eds. C. Dewitt & B. S. Dewitt (New York: Gordon and Breach), 343–450

Nulsen, P. E. J., Jones, C., Forman, W. R., David, L. P., McNamara, B. R., Rafferty, D. A., Bîrzan, L., & Wise, M. W. 2007, in Heating Versus Cooling in Galaxies and Clusters of Galaxies, eds. H. Böhringer, G. W. Pratt, A. Finoguenov, & P. Schuecker, (Berlin/Heidelberg: Springer) 210

O'Dea, C. P. 1998, PASP, 110, 493

Owen, F. N., Eilek, J. A., & Kassim, N. E. 2000, ApJ, 543, 611

Padovani, P., Mainieri, V., Tozzi, P., Kellermann, K. I., Fomalont, E. B., Miller, N., Rosati, P., & Shaver, P. 2009, ApJ, 694, 235

Perley, R. A., Dreher, J. W., & Cowan, J. J. 1984, ApJL, 285, L35

Peterson, J. R., Kahn, S. M., Paerels, F. B. S., Kaastra, J. S., Tamura, T., Bleeker, J. A. M., Ferrigno, C., & Jernigan, J. G. 2003, ApJ, 590, 207

Peterson, B. M., et al. 2004, ApJ, 613, 682

Ponman, T. J., Sanderson, A. J. R., & Finoguenov, A. 2003, MNRAS, 343, 331

Puchwein, E., Sijacki, D., & Springel, V. 2008, ApJL, 687, L53

Rafferty, D. A., McNamara, B. R., Nulsen, P. E. J., & Wise, M. W. 2006, ApJ, 652, 216

Reynolds, C. S., Heinz, S., & Begelman, M. C. 2001, ApJL, 549, L179

Richards, G. T., et al. 2006, AJ, 131, 2766

Ryle, M., & Clarke, R. W. 1961, MNRAS, 122, 349

Ryle, M., & Scheuer, P. A. G. 1955, R. Soc. Lond. Proc. A, 230, 448

Salviander, S., Shields, G. A., Gebhardt, K., & Bonning, E. W. 2007, ApJ, 662, 131

Sandage, A. 1965, ApJ, 141, 1560

Sazonov, S. Y., Ostriker, J. P., Ciotti, L., & Sunyaev, R. A. 2005, MNRAS, 358, 168

Schawinski, K., Virani, S., Simmons, B., Urry, C. M., Treister, E., Kaviraj, S., & Kushkuley, B. 2009, ApJL, 692, L19

Schmidt, M. 1963, Nature, 197, 1040

Schmidt, M., & Green, R. F. 1983, ApJ, 269, 352

Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337

Shankar, F. 2009, New Astron. Rev., 53, 57

Shapiro, S. L. 2005, ApJ, 620, 59

Short, C. J., & Thomas, P. A. 2009, ApJ, 704, 915

Siemiginowska, A., Burke, D. J., Aldcroft, T. L., Worrall, D. M., Allen, S., Bechtold, J., Clarke, T., & Cheung, C. C. 2010, ApJ, 722, 102

Silk, J., & Rees, M. J. 1998, A&A, 331, L1

Simpson, C. 2005, MNRAS, 360, 565

Smolčić, V., et al. 2009, ApJ, 696, 24

Soltan, A. 1982, MNRAS, 200, 115

Springel, V., et al. 2005, Nature, 435, 629

Steffen, A. T., Barger, A. J., Cowie, L. L., Mushotzky, R. F., & Yang, Y. 2003, ApJL, 596, L23

Steffen, A. T., Strateva, I., Brandt, W. N., Alexander, D. M., Koekemoer, A. M., Lehmer, B. D., Schneider, D. P., & Vignali, C. 2006, AJ, 131, 2826

Stern, D., et al. 2005, ApJ, 631, 163

Strateva, I., et al. 2001, AJ, 122, 1861

Tabor, G., & Binney, J. 1993, MNRAS, 263, 323

Thorne, K. S. 1994, Black Holes and Time Warps: Einstein's Outrageous Legacy, ed. K. S. Thorne (New York: W.W. Norton)

Treister, E., et al. 2006, ApJ, 640, 603

Treister, E., Urry, C. M., & Virani, S. 2009, ApJ, 696, 110

Tremonti, C. A., Moustakas, J., & Diamond-Stanic, A. M. 2007, ApJL, 663, L77

Ueda, Y., Akiyama, M., Ohta, K., & Miyaji, T. 2003, ApJ, 598, 886

Urry, C. M., & Padovani, P. 1995, PASP, 107, 803

Vernaleo, J. C., & Reynolds, C. S. 2006, ApJ, 645, 83

Volonteri, M. 2010, A&AR, 18, 279

Wagner, A. Y., & Bicknell, G. V. 2011, ApJ, 728, 29

Wall, J. V., Jackson, C. A., Shaver, P. A., Hook, I. M., & Kellermann, K. I. 2005, A&A, 434, 133

Willott, C. J., Rawlings, S., Blundell, K. M., Lacy, M., & Eales, S. A. 2001, MNRAS, 322, 536

Wise, M. W., McNamara, B. R., Nulsen, P. E. J., Houck, J. C., & David, L. P. 2007, ApJ, 659, 1153

Worsley, M. A., et al. 2005, MNRAS, 357, 1281

Wyithe, J. S. B., & Loeb, A. 2003, ApJ, 595, 614

Xue, Y. Q., et al. 2011, ApJS, 195, 10

Young, M., Elvis, M., & Risaliti, G. 2009, ApJS, 183, 17

Yu, Q., & Tremaine, S. 2002, MNRAS, 335, 965

12 The Intergalactic Medium

Renyue Cen
Department of Astrophysical Sciences, Princeton University
Observatory, Peyton Hall, Princeton, NJ, USA

1 Introduction 568

2 Dark Ages: $z = 30$–1,100 568

3 Epoch of Reionization: $z = 6$–30 570

4 Growth of Modern Structures: $z = 0$–6 573
4.1 The Global History of IGM at $z = 0$–6 575
4.2 Lyα Forest 578
4.3 Lyman Limit Systems 584
4.4 Damped Lyα Systems 585
4.5 Metal-Line Systems 589
4.6 Warm-Hot Intergalactic Medium 599
4.6.1 The Missing Metals Problem at $z = 2$–3 599
4.6.2 The Missing Baryons Problem at $z = 0$ 601

Acknowledgments 605

References 605

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_12, © Springer Science+Business Media Dordrecht 2013

Abstract: The intergalactic medium has contained, and still does, most of the matter in the universe. Galaxies form out of matter that originates from the intergalactic medium. The radiation from stars in galaxies plays an essential role in writing the ionization and thermal histories of the intergalactic medium. Galaxies return matter back to the intergalactic medium in the form of galactic winds powered by stellar winds and supernova explosions that in addition transport energy and metals. The various forms of feedback exerted on the intergalactic medium by galaxies have profound effects on subsequent galaxy formation. This two-way interaction between galaxies and the intergalactic medium is the primary driver of the formation and evolution of both. This chapter synthesizes our current knowledge of this interaction, focusing mainly on the evolution of the intergalactic medium. While it covers the entire redshift range from $z = 1{,}100$ to $z = 0$, the content is heavily skewed toward lower redshift, reflecting the current state of knowledge.

Keywords: Cosmology: observations, Intergalactic medium, Large-scale structure of universe

1 Introduction

More than 99% of the space in the present-day universe is occupied by the intergalactic medium (IGM), a tenuous gas in the intergalactic space of density lower than 10^{-5} atom per cubic centimeter. Toward higher redshift, the IGM takes a still higher percentage of space, with its mean density going up as $(1 + z)^3$, where z is redshift, due to the expansion of the universe with time. The IGM provides the fuel for galaxy formation and serves as the depository for material returned from star formation. The thermal, ionization, and metal-enrichment histories of the IGM are very complex and an active research subject area in the contemporary research of formation and evolution of galaxies and IGM.

This chapter provides an up-to-date review on these histories, chronologically separated into three periods: (1) Dark Ages, $z = 30$–$1{,}000$; (2) Epoch of Reionization, $z = 6$–30; and (3) Growth of Modern Structures, $z = 0$–6. The knowledge content of the periods increases with time, a reflection of the current state of knowledge. The presented history is expected to take place in the context of the current standard cosmological constant-dominated cold dark matter model (Komatsu et al. 2011) that is widely accepted, and where direct observations are available, it is largely in agreement with them. Alternative theories will not be described in the interest of space.

2 Dark Ages: $z = 30$–$1{,}100$

The universe is believed to begin with a hot Big Bang (Lemaître 1931). Subsequently, it expands and its content cools. By the time when the universe is about 300,000 years old, it has cooled to a temperature of about 3,000 K that hydrogen atoms begin to form, a landmark phase transition in the IGM that is termed "cosmological recombination" – the IGM switches from a plasma to a neutral medium. Most of the early history of the IGM preceding the dark ages – at $z \geq 1{,}100$ – is relatively well understood and its dynamics is analytically tractable (Peebles 1993). The state of the IGM at the cosmological recombination at $z \sim 1{,}100$ has been directly measured by the cosmic microwave background (CMB) observations, as shown in ➲ *Fig. 12-1*. At that moment, the density and temperature fluctuations in the IGM are about ten parts in a million

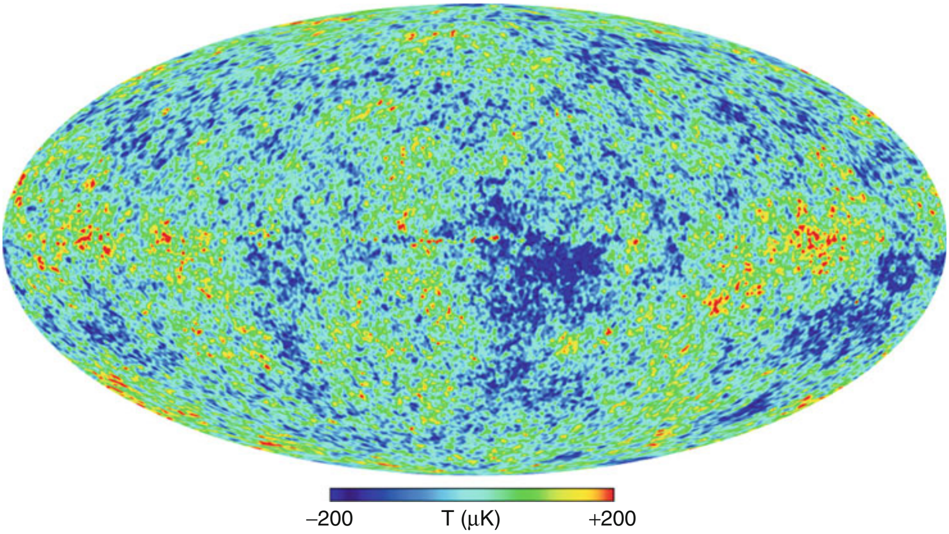


Fig. 12-1
A CMB temperature fluctuation map from the Wilkinson Microwave Anisotropy Probe (WMAP) observations (This figure is taken from Bennett et al. (2003))

(e.g., Spergel et al. 2003). These small density fluctuations will grow with time under the attractive action of gravity, in conjunction with other astrophysical processes, to ultimately produce the complex universe we see around us that is filled with galaxies, stars, and planets.

What follows is the period termed the cosmic "Dark Ages" (DA) (Peacock 1992), over which the universe grows by a factor of about 30. During this period, the IGM continues to cool passively as the universe continues to expand, until the formation of the very first gravitationally bound astrophysical system – a star or a black hole that shines, ending the DA. When the first star or black hole forms is unknown. The relevant cosmological parameters that determine the redshift of the formation of first star or black hole are the amplitude and shape of the density fluctuation power spectrum of matter. Because the formation (i.e., gravitational collapse) epoch of the very first cosmic structure depends rather sensitively on these two parameters, the exact ending redshift of the DA is imprecisely known theoretically in the context of the parameters that are presently observationally measured to an accuracy of 5–10% (Seljak et al. 2005; Komatsu et al. 2011). The most likely range is $z = 30–50$ (Barkana and Loeb 2001), when the universe is about 50–100 million years old.

The DA is thus rather uneventful: there is no star and black hole or any significant astronomical systems that are capable of producing a significant amount of energy, either tapping into gravitational or nuclear potential, to affect their surroundings. Nonetheless, four points may be noted about the IGM during the DA. First, the temperature of the IGM decreases as $(1 + z)^2$ nearly uniformly. Second, the density of the IGM decreases as $(1 + z)^3$ nearly uniformly. Third, near the end of this period some overdense regions on various small mass scales of $\leq 10^5\ M_\odot$ gradually decouple from the universal expansion and turnaround en route to gravitational collapse. Finally, the IGM is predominantly neutral with an ionization fraction of $\sim 10^{-4}$, a remnant of the cosmological recombination event. As a result, the universe is opaque to all light blueward of the Lyman alpha line wavelength of 1, 216 Å.

3 Epoch of Reionization: $z = 6$–30

Following the formation of the first star or black hole, the era of the DA comes to an end. So does the phase of passive cooling of the IGM, because stars and black holes exert feedback on the IGM in several ways. First, radiation from stars and black holes provides a source of heating and ionization of the IGM. Second, mechanical energy in the forms of winds from stars and black holes affects their surroundings, including the IGM. Finally, these winds also carry metals produced in stars, and they contaminate the primordial IGM with metals that have profound effects on subsequent formation of stars, black holes, and galaxies.

The subsequent thermal history of the IGM is determined by an intricate balance of primarily three competing processes – the universal expansion that cools it, radiative heating by stars and black holes, and gravitational interactions. This rather prolonged epoch, from the end of the DA to the end of reionization, is called the Epoch of Reionization (EoR). The prevailing theory is that the primary ionizing source for EoR are stars. While the precise history of the EoR is not known, one may make a plausible theoretical estimate. *Figure 12-2* shows the density fluctuation history and hence formation redshift of dark matter halos. In the standard cold dark matter model, most of the nonrelativistic dynamical matter is in an unknown

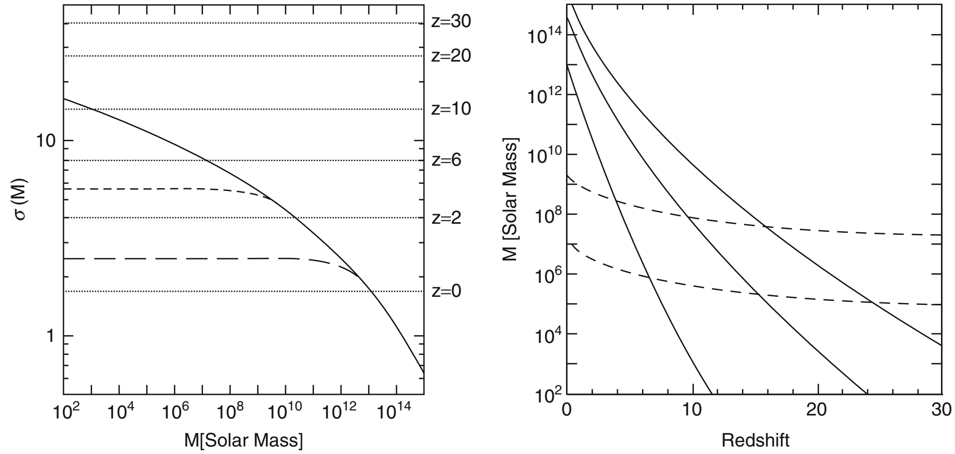


Fig. 12-2

Left panel: mass fluctuations and collapse thresholds in cold dark matter models. The _horizontal dotted lines_ show the value of the extrapolated collapse overdensity $\delta_{\text{crit}}(z)$ at the indicated redshifts. Also shown is the value of the variance of density fluctuations $\sigma(M)$ for a cold dark matter model as well as $\sigma(M)$ for a power spectrum with a cutoff below a mass $M = 1.7 \times 10^8\ M_\odot$ (_short-dashed curve_), or $M = 1.7 \times 10^{11}\ M_\odot$ (_long-dashed curve_). The intersection of the horizontal lines with the other curves indicate, at each redshift z, the mass scale (for each model) at which a 1σ fluctuation is just collapsing at z. _Right panel_: characteristic properties of collapsing halos–halo mass. The _solid curves_ show the mass of collapsing halos which correspond to 1σ, 2σ, and 3σ fluctuations (in order from bottom to top). The _dashed curves_ show the mass corresponding to the minimum temperature required for efficient cooling with primordial atomic species only (_upper curve_) or with the addition of molecular hydrogen (_lower curve_) (This plot is taken from Barkana and Loeb (2001) where the reader can also find the relevant cosmological parameters used for it)

form dubbed "cold dark matter" that is believed to be nonbaryonic. The dynamics on galactic and larger scales are largely determined by dark matter halos, within which galaxies form and live. Equipped with this basic knowledge and plausible (but very uncertain) assumptions about the astrophysics of galaxy formation in that era, one can sketch out the history of IGM in the EoR era. *Figure 12-3* shows a theory where the universe turns out to be reionized twice, once at $z \sim 15$ and then again at $z \sim 6$, in the context of the cold dark matter model (Cen 2003). This model is consistent with all extant observations, including quasar absorption spectrum measurements at $z \sim 6$ (Fan et al. 2006) and the CMB observations (Komatsu et al. 2011). The bimodal reionization picture originates from the bimodal nature of stellar radiation efficiency, where Population III (Pop III) stars – the very first generation of stars that form out of prestine primordial gas – are much more efficient, by a factor of ~20, in producing hydrogen ionizing photons than the subsequent Pop II stars that form out of metal-enriched gas (Bromm et al. 2001). In the left panel, the inflection point at $z \sim 30$ follows the birth of the very first star at a slightly higher redshift. A sustained ascent in the mean temperature of the IGM from $z \sim 30$ up to the redshift of the first reionization at $z \sim 15$ reflects a continued heating by an increasing star formation rate with time. Subsequently, the mean temperature of the IGM is roughly maintained at 10^4 K, due to the action of two counter-balancing effects: substantial cooling reduces the Jeans mass of the IGM and increases the star formation rate, which then provides increased heating by photoionization that in turn suppresses star formation. Thus, the mean temperature of the IGM is self-regulated due to the competition between the cooling (mostly Compton cooling by the cosmic microwave background) and photoheating due to ionization of hydrogen and helium atoms by Lyman continuum ultraviolet photons. In the cold dark matter cosmological model with the measured parameters (Komatsu et al. 2011), it happens that the star formation

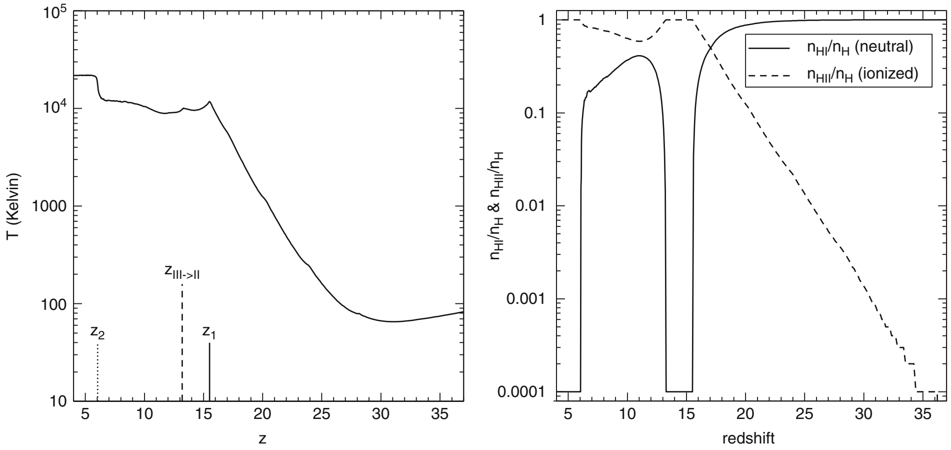


Fig. 12-3

***Left panel*: shows the evolution of the mean IGM temperature as a function of redshift during the second cosmological reionization process. The *solid vertical tick* indicates the first reionization epoch. The *dashed vertical tick* indicates the transition epoch from Pop III stars to Pop II stars. The *dotted vertical tick* indicates the second reionization epoch. *Right panel*: shows the global mean of the hydrogen neutral (*solid*) and complimentary ionized (*dashed*) fraction as a function of redshift (This figure is taken from Cen (2003))**

activities are at a level such that the two balancing terms are comparable in magnitude when temperature is maintained at ~10^4 K, thus resulting in a fairly mild evolution of the IGM mean temperature seen. By redshift $z \sim 6$, the star formation rate has surpassed a threshold such that the rate of recombination between protons and electrons in a significantly neutral medium is no longer capable of balancing the photo-ionization rate by stars: the universe is now fully ionized finally and for all.

The temperature evolution shown in the left panel of *Fig. 12-3* is accompanied by the global evolution of ionization fraction shown in the right panel. The first reionization at $z \sim 15$ as well as the sustained ionized state until $z \sim 13$ is made possible by Pop III stars. The redshift $z \sim 13$ marks the transition from Pop III stars to Pop II stars, which occurs after a fraction 10^{-4} of total gas has formed into Pop III stars and IGM is metal-enriched to a critical value, causing the emission of hydrogen ionizing photons to plunge. The suddenly reduced hydrogen ionizing photon emission is no longer able to counter the rapid hydrogen recombination process, resulting in the second cosmological recombination at $z \sim 13$. Since a very small amount of neutral fraction suffices to blank out all Lyα emission, the universe essentially becomes opaque to Lyα photons from $z \sim 13$ to $z \sim 6$.

A key issue is the ionizing photon budgetary constraints during cosmological reionization: how many ionizing photons per atom are required? Since the IGM contains nearly all the baryons in the universe during EoR, the minimum, necessary requirement is one ionizing photon per hydrogen atom. To facilitate understanding, it is useful to show some important time scales involved. The left panel of *Fig. 12-4* shows the ratio of hydrogen recombination time over the Hubble time and the ratio of Compton cooling time over the Hubble time. It is noted

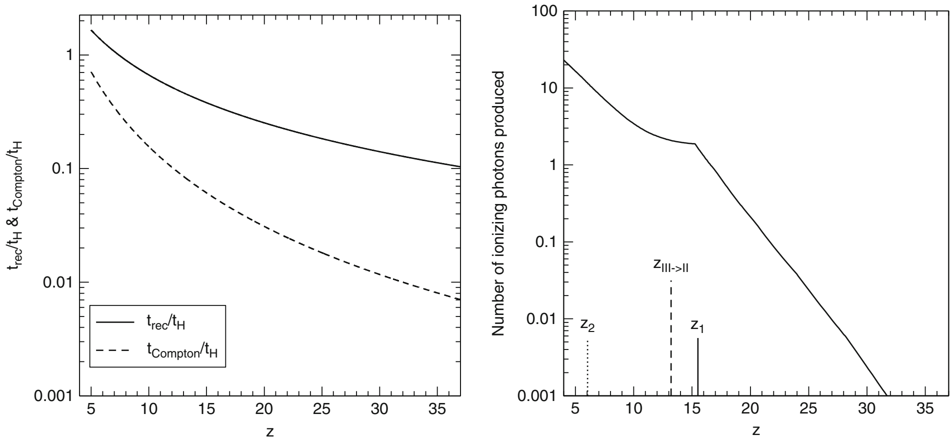


Fig. 12-4

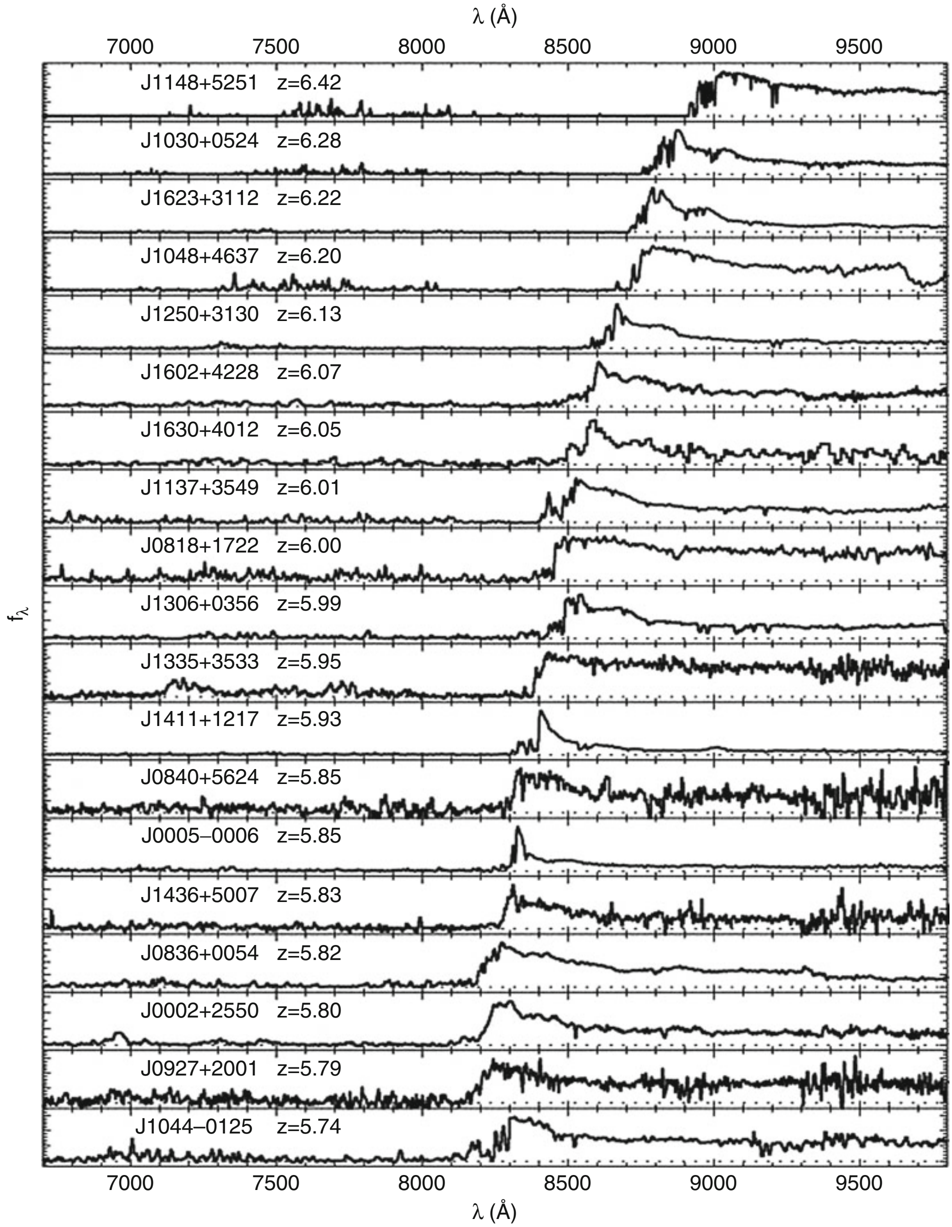
***Left panel*: shows the ratio of the recombination time to the Hubble time (*solid curve*) and ratio of the Compton cooling time to the Hubble time (*dashed curve*), a function of redshift. *Right panel*: shows the cumulative number of ionizing photons per hydrogen atom produced as a function of redshift. The *solid vertical tick* indicates the first reionization epoch. The *dashed vertical tick* indicates the transition epoch from Pop III stars to Pop II stars. The *dotted vertical tick* indicates the second reionization epoch (This figure is taken from Cen (2003))**

that at the redshift and density of interest, Compton cooling dominates over other cooling processes for the general IGM, although adiabatic cooling starts to become important approaching the end of the final reionization epoch. It is seen that the hydrogen recombination time is significantly longer than the Compton cooling time, both of which are, however, shorter than the Hubble time at $z \geq 8$. Therefore, the IGM at early times heated up by the photoionization would subsequently not only cool down but also recombine. Because ionized hydrogen has a tendency to recombine to become neutral and because the recombination time scale is shorter than the Hubble time, the actual number of ionizing photons per hydrogen needs to exceed unity in order to fully reionize the universe. From the right panel of ➋ *Fig. 12-4*, it is seen that at the epoch of the first reionization about two ionizing photons per baryon have been produced by Pop III stars. About 10^{-4} fraction of gas formed into Pop III stars then. More than ten ionizing photons per baryon have been produced by the time the universe is reionized for the second time at $z = 6$. By this time, about 1% of the total gas has formed into Pop II stars that are primarily responsible for the final reionization. The more stringent requirement for ionizing photons at $z \sim 6$ than at $z \sim 15$ stems from the fact that the IGM becomes increasingly clumpy in nature, which substantially lowers the recombination time scale of the IGM as a whole, even though the mean density of the IGM has decreased with time.

The starting redshift as well as the overall evolution of the IGM in the EoR era are still quite uncertain. Other plausible models can be designed to meet extant observational constraints (e.g., Haiman and Holder 2003). The primary constraint on the duration and starting redshift comes from the Wilkinson Microwave Anisotropy Probe (WMAP) observations that give an integral constraint of the Thomson optical depth $\tau_e = 0.087 \pm 0.017$ due to scattering by free electrons in the IGM (e.g., Dunkley et al. 2009). The above double reionization picture yields $\tau_e = 0.10 \pm 0.02$ that is consistent with this constraint. The ending redshift of EoR, if defined as the redshift when the universe becomes transparent to Lyα photons, is, however, very well constrained observationally by quasar absorption spectra of the hydrogen Lyman alpha from the Sloan Digital Sky Survey (SDSS) (Fan et al. 2006), as shown in ➋ *Fig. 12-5*. It is seen in ➋ *Fig. 12-5* that there is a complete lack of photon flux down to an observed wavelength of about $\lambda_{\rm obs} \approx 8,300$ Å, whereas measurable non-zero flux is detected at still shorter wavelengths for each spectrum of the high-redshift quasars. The observed wavelength $\lambda_{\rm obs}$ is related to the intrinsic restframe wavelength $\lambda_{\rm int} = 1,216$ Å by $\lambda_{\rm obs} = 1,216(1+z)$ Å. There is clear evidence that redshift $z \sim 5.8$ marks a transition from an opaque to transparent universe to Lyα photons, another landmark event in the cosmic history.

4 Growth of Modern Structures: $z = 0$–6

Following the completion of the cosmological reionization at $z \sim 6$, the IGM becomes increasingly more transparent, primarily due to the fact that the mean optical depth of the IGM at the restframe Lyα wavelength decreases with redshift approximately as $(1+z)^{3/2}$ at a fixed neutral fraction. The number density of quasars, which are the background light sources for absorption spectrum observations, is observed to increase rapidly with decreasing redshift. In combination, the IGM can be probed with increasing precision and statistics at lower redshift. The amount of direct observational data is very rich. Because of this reason, this section is subdivided into several subsections organized around distinct IGM absorption systems. First in ➋ Sect. 4.1 a description of the global evolution of the IGM from $z \sim 6$ to $z = 0$ is given, followed by



■ Fig. 12-5

Spectra of a sample of 19 SDSS quasars at $5.74 < z < 6.42$. Twelve of the spectra were taken with Keck ESI, while the others were observed with the MMT Red Channel and Kitt Peak 4 m MARS spectrographs (This figure is taken from Fan et al. (2006))

five sections devoted to five different topics, Lyα forest (➜ Sect. 4.2), Lyman limit systems (➜ Sect. 4.3), damped Lyman alpha systems (➜ Sect. 4.4), metal-line systems (➜ Sect. 4.5), and the warm-hot intergalactic medium (➜ Sect. 4.6).

4.1 The Global History of IGM at $z = 0$–6

Before describing more quantitatively, it is useful to give a simple and surprisingly accurate order of magnitude physical argument. In all standard pictures for the growth of structure, there is imprinted at an early time a spectrum of perturbations, with the amplitude of the fluctuations being largest at small scales and smaller at larger scales. After decoupling at $z \sim 1,100$, all waves grow (due to self-gravity) roughly as $(1+z)^{-1}$. They have reached the nonlinear length and mass scales of 5–7 h^{-1} Mpc and (5–10) × 10^{13} h^{-1} $M_\odot$, respectively, by $z = 0$. When a perturbation of a given scale L collapses at time t due to gravity, geometry indicates it must do so with a velocity $v \approx L/t$ and, as opposite sides of the collapsing perturbation meet and attempt to cross one another, a shock is generated behind which the sound speed is $C \approx v$. Combining these simple arguments and noting that $t^{-1} \approx H_z$ gives us

$$C_z^2 = K(H_z L_{nl})_z^2, \qquad (12.1)$$

where K is a numerical constant and H_z and $L_{nl,z}$ are the Hubble constant and nonlinear length scale at epoch z, respectively. Applying this to the current epoch (with $K = 1$) gives $C_o \sim 500$–700 km s^{-1}, which corresponds to a temperature of (2–4) × 10^7 K. This should correspond to the typical temperature of recently collapsed (i.e., high density) objects, with the global mean temperature somewhat lower. Note that K is only a dimensionless number to indicate the nature of the scaling relation and is not intended to model shock jump conditions.

Detailed cosmological simulations produce results that are consistent with this physical argument. ➜ *Figure 12-6* shows the mean computed temperature as a function of redshift based on cosmological simulations. The volume weighted average temperature at $z = 0$ is $10^{5.5}$ K, the mass (ρ) weighted average temperature is $10^{7.4}$ K (the value given by (➜ 12.1) with $K = 1$), and the ρ^2 weighted average (roughly speaking emission weighted) is $10^{8.0}$ K. Thus, the high-density regions are in just the temperature domain indicated by (➜ 12.4) with the constant taken to be unity; in other words, (➜ 12.1) (adopting $K=1$) in fact roughly tracks the density-weighted averages from redshift three onward, confirming the simple physical picture presented earlier.

Now examine the overall thermal evolution of the IGM in greater detail, dividing the IGM into three temperature ranges: (1) $T < 10^5$ K cold-warm gas, a dominant portion of which is the Lyα forest described in ➜ Sect. 4.2; (2) 10^7 K> $T > 10^5$ K gas, which is termed "warm-hot intergalactic medium" (WHIM) and is mainly in unvirialized regions that will be detailed in ➜ Sect. 4.6; and (3) $T > 10^7$ K, the normal X-ray emitting gas, predominantly in collapsed and virialized X-ray clusters of galaxies. A last component (4) is the cold gas that has been condensed into stellar objects, which is designated as "galaxies." ➜ *Figure 12-7* shows the evolution of these four components. The overall evolution of the four components are in good agreement and relevant observations (e.g., Fukugita et al. 1998). In particular, the hot component ($T > 10^7$ K) increases in mass fraction, reaching 12% by mass at $z = 0$, and is consistent with observations of the local properties of the X-ray emitting great clusters of galaxies. The condensed component remains small (8%) but consistent with the observed mass density in galaxies at $z = 0$ (Cole et al. 2001). While most of the volume is always filled with $T < 10^5$ K gas (Lyα forest) at all

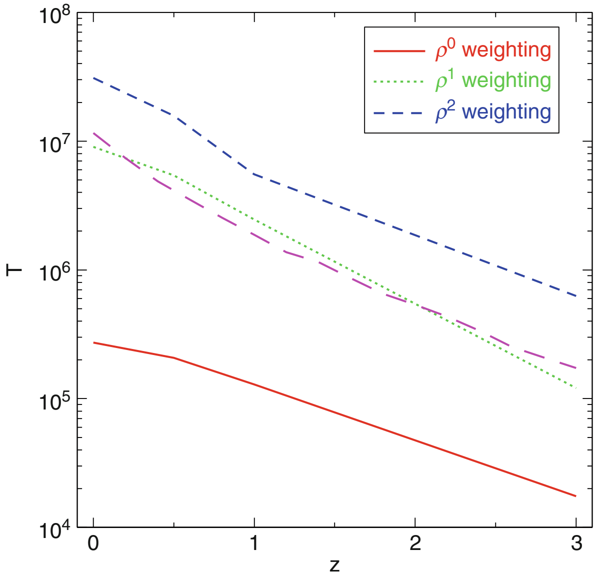


Fig. 12-6
Shows globally averaged temperatures within our 100 h^{-1} Mpc^3 box as a function of redshift from the simulation. The *solid*, *dotted*, and *short-dashed curves* are the average temperatures weighted by volume, density, and density squared, respectively. The *long-dashed curve* represents the results from the simple physical argument as indicated by (➋ 12.1) with the constant *K* = 1 and indicates that the temperature of the high-density nonlinear regions is well represented by that formula (This figure is taken from Cen and Ostriker (1999))

redshift, the mass fraction of the cold-warm IGM declines from nearly 100% at $z > 4$ to 40% at $z = 0$, consistent with both high (Rauch 1998) and low redshift observations (Penton et al. 2004) of Lyα forest. About 40–50% of all baryons are in WHIM by the time $z = 0$.

Each of the IGM components is composed of different regions that have gone through distinct evolutionary paths and thus spans a wide range in density, as shown in ➋ *Fig. 12-8*. The distribution of the cold-warm component (triangles) is always peaked at the mean density at all redshifts, reflecting the initial Gaussian distribution of gas around the cosmic mean and indicating that the bulk of the IGM at mean density or lower has never been shock heated by either gravitational shocks or feedback shocks due to galactic winds. The cold-warm gas extends to very high densities ($\geq 10^5$). It is interesting to note that the amount of cold-warm gas that could potentially feed the star formation, i.e., the cold-warm gas at density $\log \rho / < \rho > \geq 2$–3, remains constant, within a factor of ~2, over the range redshift shown $z = 0$–5. This is consistent with observations of the nearly nonevolving amount of gas probed by DLAs (e.g., Péroux et al. 2003). The physical relation between this apparently nonevolving gas and the precipitous drop of star formation rate at $z < 1$ is currently unclear.

The distribution of the WHIM also appears to peak at a constant overdensity of about ten times the mean density. This is rather intriguing. In order to properly interpret this interesting phenomenon, it is useful to understand the heating sources of WHIM. There are two primary

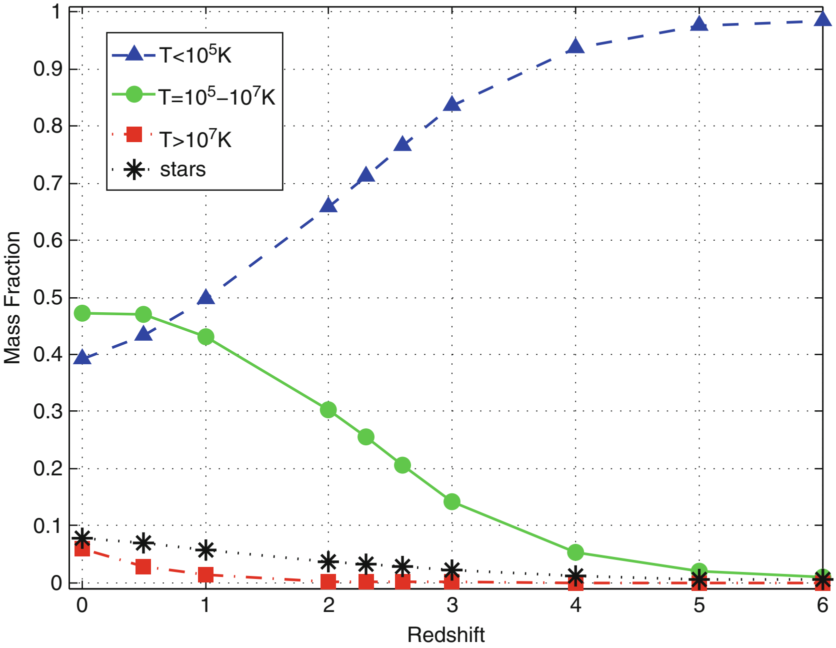


Fig. 12-7

Shows the evolution of baryons for the four mutually exclusive components: (1) $T < 10^5$ K cold-warm gas, (2) WHIM at 10^7 K > T > 10^5 K, (3) Hot X-ray emitting gas at $T > 10^7$ K and (4) "stars" (This figure is taken from Cen and Chisari (2011))

heating sources for WHIM: shocks due to the collapse of large-scale structure and galactic wind-produced shocks. Studies have shown that gravitational shock heating due to the formation of large-scale structure dominates the energy input for heating up and thus turning about 50% of the IGM into WHIM by $z = 0$ (Cen and Ostriker 1999; Davé et al. 2001; Cen and Ostriker 2006). It is, however, expected that heating due to hydrodynamic shocks emanating from galactic winds become increasingly more important at higher redshifts because the amount of energy from gravitational collapse of large-scale structure as well as the resulting shock velocity decreases steeply toward higher redshift (Cen and Chisari 2011). For the case of galactic wind-generated shocks, if gas is shock heated to 10^5 K, the shock velocity is roughly 70 km s^{-1}. With that velocity, the shock will be able to travel roughly $700(1+z)^{-1}$ kpc comoving over the Hubble time at any redshift. Therefore, one should expect to see shocks have reached a few hundred kpc comoving at any redshift, which are about one to a few times the virial radius of typical large galaxies, which in turn correspond to an overdensity in the vicinity of 10 and are thus in good agreement with simulation results. Some shocks penetrate deeper into the IGM, especially along directions with lower densities and steeper density gradients, but the amount of mass effected in these low-density regions is small, corresponding to the sharp drop of WHIM mass at the low-density end (➋ *Fig. 12-8*). This last point is best corroborated by the distribution of the hot gas at high redshift ($z = 3, 5$), in the bottom two panels of ➋ *Fig. 12-8*. There, a small amount of hot gas heated up by GSW shocks is indeed produced in regions of density lower

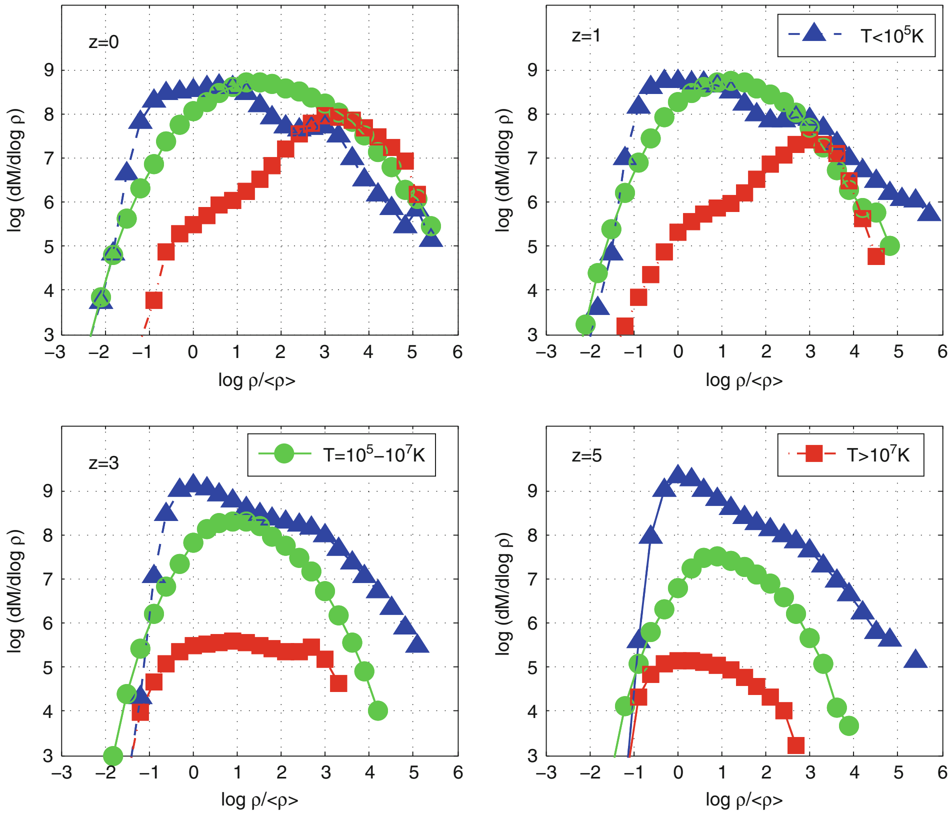


Fig. 12-8
Shows the mass distribution of the three IGM components – (1) cold-warm gas at $T < 10^5$ K, (2) WHIM at 10^7 K > $T > 10^5$ K, (3) Hot X-ray emitting gas at $T > 10^7$ K – as a function of overdensity at four different redshifts $z = 0, 1, 3, 5$. Note that the area under each curve is proportional to the mass contained (This figure is taken from Cen and Chisari (2011))

than the mean density and traces a larger amount of WHIM gas that is also produced there. At $z = 1$ and lower, some comparable, small amount of hot gas is still produced at low-density regions. But the vast majority of hot X-ray emitting gas is now residing in the deep potential wells of X-ray clusters of galaxies, when the cluster scale turns nonlinear and collapses.

4.2 Lyα Forest

The previous section shows that most of the IGM at moderate to high redshift ($z > 1$) should be in the cold-warm phase that comprises mostly the Lyα forest. For a background quasar its emitted light redshifts to longer wavelength, when traveling toward the observer on Earth. For a photon of wavelength shorter than the Lyα line wavelength, it may be redshifted to the restframe Lyα line center at some redshift. If at that redshift a significant amount of neutral hydrogen atoms intercept the line of sight to the quasar, that photon will be scattered out of the line of

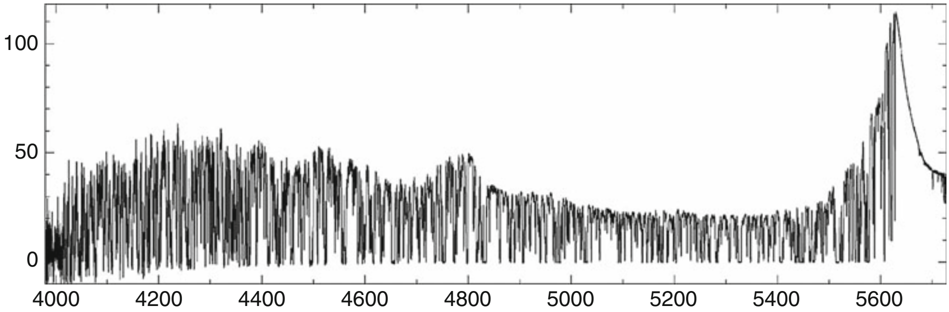


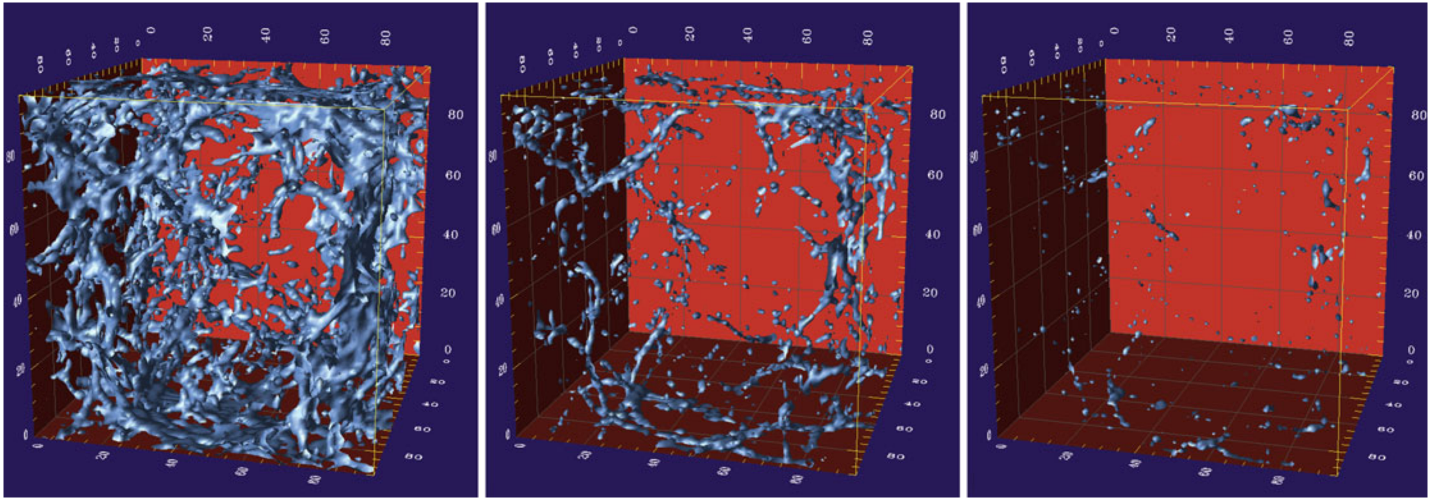
Fig. 12-9
High-resolution (full width at half maximum *FWHM* ≈ 6.6 km s^{-1}) spectrum of the $z_{\rm em}$ = 3.62 QSO1422+23 (V = 16.5), taken with the Keck High Resolution Spectrograph (HIRES) (signal-to-noise ratio 150 per resolution element, exposure time 25,000 s) (Data from Womble et al. (1996))

sight, causing a dip at the corresponding wavelength in the observed flux spectrum. This could occur for photons of different initial wavelengths, resulting in a forest-like quasar spectrum, hence the Lyα forest. *Figure 12-9* shows an example of an observed quasar spectrum. One can measure the corresponding neutral hydrogen column density for each spectral dip. The neutral hydrogen column density for Lyα forest ranges from 10^{12} to 10^{17} cm^{-2}. Because the background quasar and its foreground absorbers are unrelated, absorption spectrum provides an unbiased probe of the IGM.

The physical origin of neutral hydrogen concentrations that produce the spectral dips in the quasars spectrum was not fully understood until mid-1990s, when cosmological hydrodynamic simulations show that the fluctuating density field at moderate redshift during structure formation can naturally explain the observed Lyα forest (Cen et al. 1994; Zhang et al. 1995; Hernquist et al. 1996; Miralda-Escude et al. 1996). *Figure 12-10* shows the iso-density surface structure of the IGM at three different densities. It is seen that the IGM forms an interconnected network of sheets and filaments at 3 times the mean density (or below) at $z = 3$, whereas at 30 times the mean density the regions are well separated and isolated. Roughly speaking, at $(3, 10, 30)$ times the mean density, the corresponding column density of Lyα clouds is $(10^{13} - 10^{14}, 10^{14} - 10^{15}, 10^{15} - 10^{16})$ cm^{-2}, respectively, at redshift $z = 2$–4.

Figure 12-11 shows detailed distribution of density, temperature, and velocity divergence in a thin slice of IGM at $z = 3$. Placing a quasar at the left end of the top horizontal dashed line (out of the four horizontal lines) one obtains an example synthetic spectrum shown in the top panel of *Fig. 12-12* along with the physical properties of the IGM that gives rise to the obtained spectral features. From *Fig. 12-11* to *Fig. 12-12* it is clear that intersections of lines of sight to quasars with the cosmic web produce the most common Lyα forest lines. These regions tend to be moderately overdense and demarcated by outward shocks at both sides. An ensemble of such synthetic spectra may be generated and observables extracted to be compared with direct observations. It is found that the standard cold dark matter model can reproduce the observed Lyα forest without introducing any additional free parameters. *Figure 12-13* shows one measure of such agreements between theory and observations with respect to the distribution of column densities of Lyα forest lines.

The general excellent agreement between observations and standard cold dark matter model, in terms of statistics including flux distribution, column density distribution, and



■ Fig. 12-10
Three-dimensional isodensity surfaces for densities at 3 (*left panel*), 10 (*middle panel*), and 30 (*right panel*) times the mean baryon density of the universe at redshift $z = 3$ (This figure is taken from Cen and Simcoe (1997))

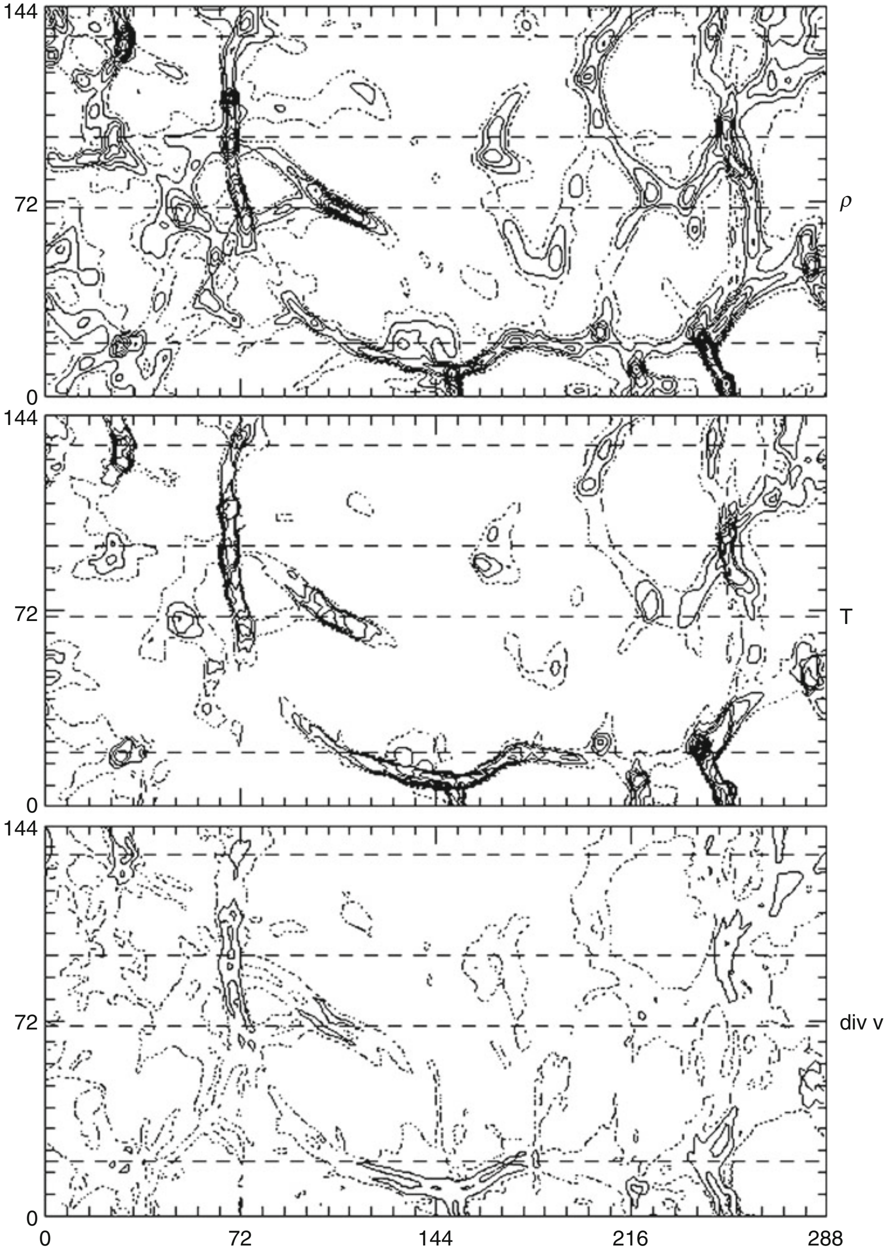


Fig. 12-11
Top panel: gas density with contours at $\rho/\overline{\rho} = 10^{0.25(i-1)}$, $i = 1, 2, 3, \ldots$ for solid contours and $i = 0$ for dotted contour. *Middle panel*: gas temperature contours at $10^{4.2+0.1i}$ K, $i = 1, 2, 3, \ldots$ for solid contours and $i = 0$ for dotted contour. *Bottom panel*: peculiar velocity divergence contours at $\nabla \cdot \mathbf{v}_p = -3H$ (*solid contour*, corresponding to constant proper density) and $\nabla \cdot \mathbf{v}_p = 0$ (*dotted contour*, corresponding to constant comoving density) (This figure is taken from Miralda-Escude et al. (1996))

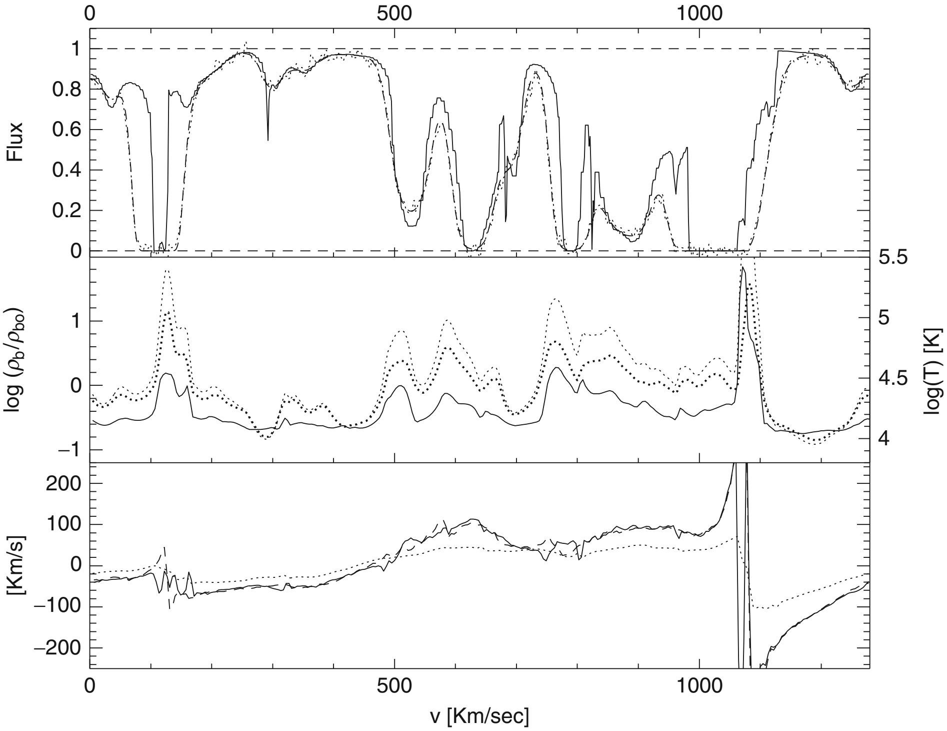


Fig. 12-12
Middle panel **in each figure shows the gas density along a row, in units of the average gas density (*thick dotted line*; *left vertical axis*); the gas temperature (*solid line*; *right vertical axis*); and the gas pressure (*thin dotted line*; the same scale as density but arbitrary units). Spatial coordinate in horizontal axis is $x = v/H$. Rows shown in *Fig. 12-7*a, b, c, d are marked as *dashed lines* in slice in *Fig. 12-5*. Calculated Lyα absorption spectrum is in *top panel*, without thermal broadening (*solid line*), and including it (*dashed line*). Peculiar velocity is shown as *dotted line* in *bottom panel*, together with gravitational acceleration (*dashed line*) and total acceleration (*solid line*) divided by the Hubble constant (This figure is taken from Miralda-Escude et al. (1996))**

Doppler width distribution, allows one, in turn, to use Lyα observations to precisely constrain cosmological parameters. CMB observations provide tight constraints on many fundamental cosmological parameters including the power spectrum of the matter density fluctuation, but CMB observations alone do not have the necessary leverage to precisely constrain the slope of the power spectrum. The Lyα forest flux distribution provides the only competitively accurate measurement of the matter power spectrum at small scales. The statistical accuracies to determine the amplitude, slope, and curvature of the density power spectrum using Lyα forest from large SDSS QSO samples have reached unprecedented 1–3% level (e.g., Mandelbaum et al. 2003). Thus, Lyα forest observations and CMB observations jointly will be able to nail down the matter power spectrum to a unprecedented level that may test fundamental theories, such as inflationary theories (Seljak et al. 2005).

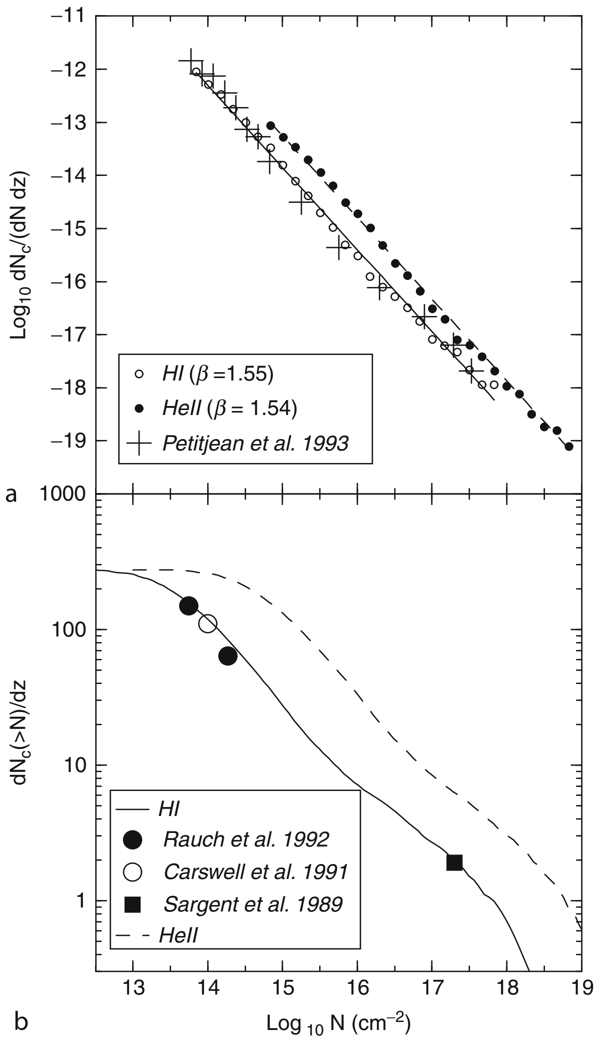


Fig. 12-13

Column density distribution per unit redshift at $z = 3$. (a) Distribution per unit column density: the *open circles* are our numerical data for HI, the *solid line* is a power-law fit with exponent $\beta = 1.55$, and the crosses are observed data. The *filled circles* are out numerical data for HeII, and the *dashed line* is a power-law fit with exponent $\beta = 1.54$. Note that there is an apparent deficiency of clouds at column densities greater than 10^{15} and 10^{16} for HI and HeII, respectively. (b) Integrated distribution: the *solid line* is for HI, and the *filled circles*, *open circle*, and the *filled square* are observational data. Also shown is the column density distribution for HeII (*dashed line*) (This figure is taken from Zhang et al. (1995))

4.3 Lyman Limit Systems

Over about ten decades of cloud column density from 10^{12} to 10^{22} cm^{-2}, the interval from 10^{17} to 10^{20} cm^{-2} is termed the Lyman limit systems (LLSs), because they are optically thick to Lyman continuum photons at $\lambda < 912$ Å. Since the completion of cosmological reionization (i.e., at $z < 6$), LLSs determines how far a Lyman continuum photon can travel before it gets absorbed by a neutral hydrogen atom. As a result, LLSs are responsible for establishing the amplitude of ionizing ultraviolet (UV) background at $z < 6$ and are cosmologically important entities. While a significant sample of observed LLS is available (e.g., Storrie-Lombardi et al. 1994; Prochaska et al. 2010), their properties are not well understood for two reasons. First, observationally, LLSs have absorption profiles that are saturated (i.e., the optical depth at the core is substantially greater than unity) but are not damped (see Sect. 4.4 for damped Lyman alpha systems). Because of that, it is difficult to precisely measure the column density of an LLS, even with high-resolution high signal-to-noise spectra. This is evident in *Fig. 12-14* by the large errorbar for the column density range where LLS are located. Second, theoretically, unlike Lyα forest, the LLSs are optically thick to Lyman continuum photons. Consequently, it is much more difficult to simulate LLS, and currently no fully consistent model has been enacted to statistically address the issue, although several pioneering attempts have been made (e.g., Katz et al. 1996; Gardner et al. 1997a; Kohler and Gnedin 2007). In conclusion, while LLSs are a very important component of the Lyα absorption lines, much progress still needs to be made to have a satisfactory physical understanding.

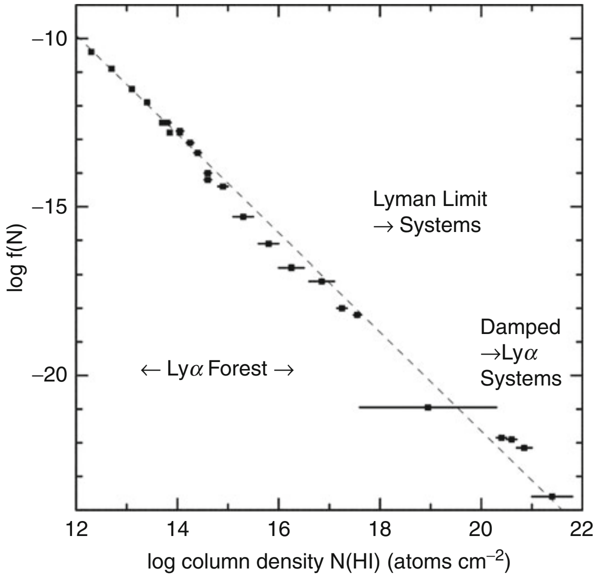


Fig. 12-14

Column density distribution function of neutral hydrogen for the 12 ≤ log N ≤ 22 Lyα absorbers. To first order, it is fitted by a power law, $f(N) \propto N^{-1.46}$, over 10 orders of magnitude in column density (This figure is taken from Storrie-Lombardi and Wolfe (2000))

4.4 Damped Lyα Systems

Damped Lyα systems (DLAs) are significant because they contain most of the neutral gas in the universe at all times since cosmological reionization (e.g., Storrie-Lombardi and Wolfe 2000; Péroux et al. 2003; Prochaska and Wolfe 2009). Molecular clouds, within which star formation takes place, condense out of cold dense neutral atomic gas contained in DLAs, evidenced by the fact that the neutral hydrogen (surface) density in DLAs and molecular hydrogen (surface) density in molecular clouds form a continuous sequence (e.g., Zwaan and Prochaska 2006), as shown in ➋ *Fig. 12-15*. Therefore, DLAs hold key to understanding the fuel for and ultimately galaxy formation. A substantial amount of recent theoretical work has been devoted to studying the nature of DLAs (e.g., Pontzen et al. 2008; Tescari et al. 2009; Hong et al. 2010) since the first investigation of Katz et al. (1996) in the context of the cold dark matter (CDM) cosmogony. Here, an up-to-date account is given based on lastest cosmological hydrodynamic simulations (Cen 2012).

Until now, no single model has been able to match the observed velocity width distribution of DLAs in that models underpredict the abundance of large velocity width DLAs, where velocity width is defined to be the velocity interval of 90% of the optical depth of the unsaturated Si II λ1808 (or other low-ionization metal lines) absorption line associated with the DLA, v_{90} (Prochaska and Wolfe 1997); the velocity structure in Lyα flux of a DLA is "damped" and does not provide the kinematic information of the underlying physical cloud. ➋ *Figure 12-16* shows the velocity width distribution at three redshifts ($z = 1.6, 3.1, 4.0$) from latest simulations. The two sets of curves, one labeled "C" and the other "V," are expected to bracket the true prediction

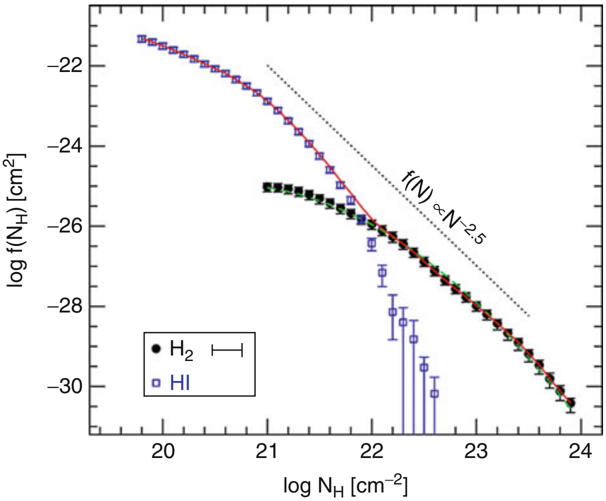


Fig. 12-15

Column density distribution function of HI and H_2 at $z = 0$. Column densities are expressed in atoms cm^{-2}, also for H_2. The *solid line* shows the summed $f(NH)$. The *horizontal error bar* indicates the uncertainty in $f(N)$ if the CO/N conversion factor is uncertain by 50%. The $f(N)$ can be fitted very well with a lognormal distribution, where $\mu = 20.6$, $\sigma = 0.65$, and the normalization is 1.1 × 10^{-25} cm^2, as indicated by the *dashed line* (This figure is taken from Zwaan and Prochaska (2006))

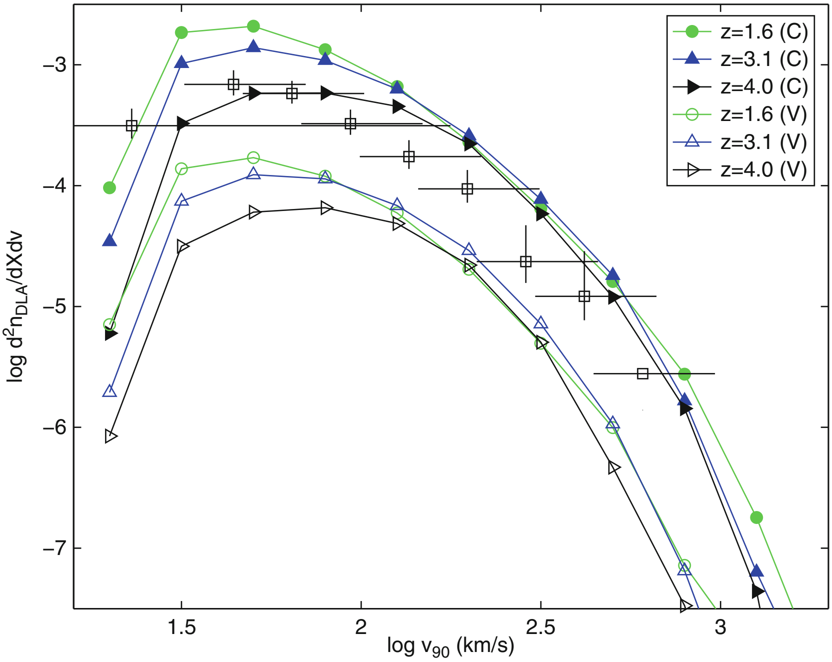


Fig. 12-16

Velocity distribution functions, defined to be the number of DLAs per unit width velocity per unit absorption length, at $z = 1.6, 3.1, 4.0$. Two sets of simulation results are shown, one for the "C" run (*solid symbols*) and "V" run (*open symbols*). The corresponding observational data for each of the individual redshifts (Prochaska et al. 2005) are shown as *open squares*, which span the redshift range of $z = 1.7$–4.5 (This figure is taken from Cen (2012))

of the cold dark matter model. Insofar as the observed velocity width distribution function lies in between the two bracketing sets and the shape of the functions are in excellent agreement with observations, including the high-velocity tail ($v_{90} \geq 300\,\mathrm{km\,s^{-1}}$), one could claim that the standard cold dark matter model can successfully explain this observed DLAs with respect to the velocity width distribution. The most important new ingredient in the latest simulations are large-scale simulations with very high resolution, properly sampling and resolving close interactions of galaxies that previously were not adequately modeled. These close interactions significantly increases DLA cross section for each galaxy in a mass-dependent fashion in that, on average, larger galaxies that tend to experience more interactions than smaller ones have larger increase in DLA cross sections relative to their stellar disk size or a fixed fraction of the area within the halo virial radius. The ubiquity of large-scale neutral cross sections at high redshift is illustrated in *Fig. 12-17*, where the DLA cross section and disk size for a random set of four galaxies are contrasted and ubiquitous extended structures are common due to galaxy-galaxy interactions at very high redshift ($z \geq 6$).

This newly found physical nature of DLAs at high redshift – extended filamentary structures instead of stellar disks – has its signature imprinted in the metallicity and size distribution of DLAs. *Figure 12-18* shows the DLA metallicity distributions at four redshift, $z = 0, 1.6, 3.1, 4.0$.

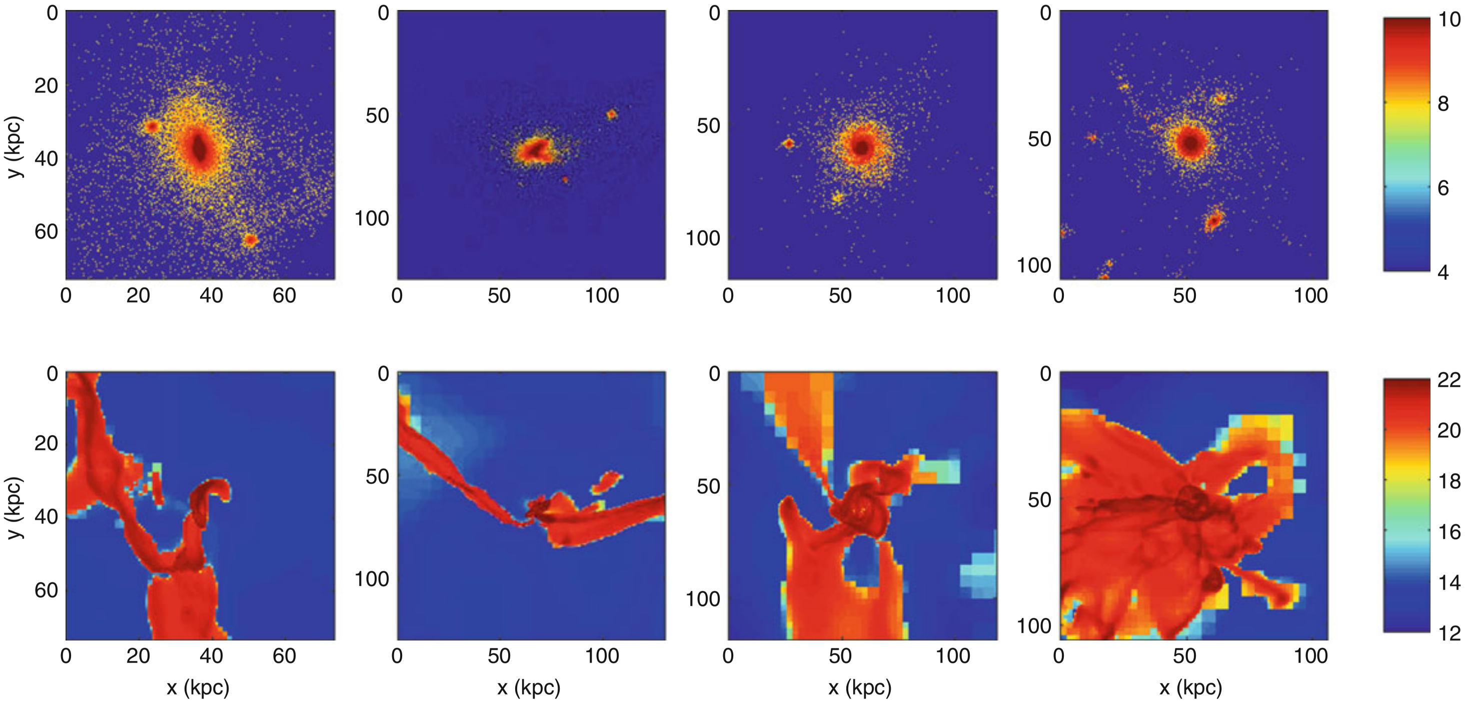


Fig. 12-17

Shows four examples of randomly chosen galaxies that have associated DLAs at $z = 3.1$. The odd rows are the logarithm of the stellar luminosity surface density maps in SDSS z band in units of $L_{\odot}$ kpc^{-2}. The even rows are the logarithm of the corresponding neutral gas column density maps in units of cm^{-2}. The lengths are in proper kpc, and the depth of each projection is about the virial diameter of each galaxy (This figure is taken from Cen (2012))

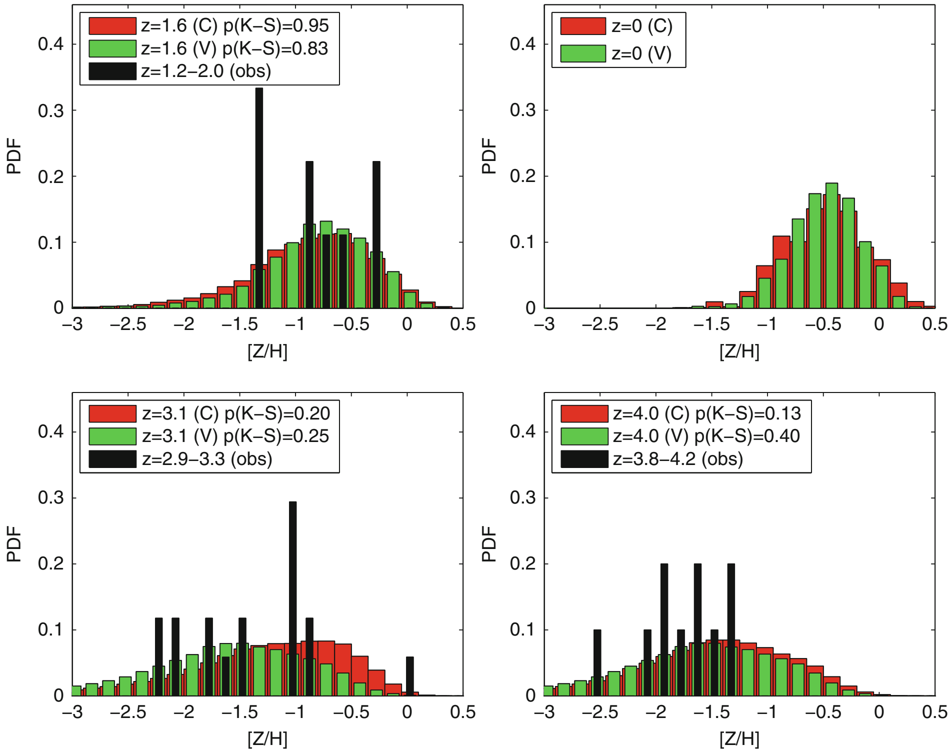


Fig. 12-18

Shows the DLA metallicity distributions at four redshift, $z = 0, 1.6, 3.1, 4.0$, for both "C" (*red histograms*) and "V" (*green histograms*) run. The observational data are from Prochaska et al. (2005), shown as *black histograms*. Because there is non-negligible evolution, the comparisons between simulations at a given redshift are only made with observed DLAs within a narrow redshift window, as shown. Probabilities that simulated and observed samples are drawn from the same underlying distribution are indicated in each panel, separately for "C" and "V" run (This figure is taken from Cen (2012))

For the three redshifts, $z = 1.6, 3.1, 4.0$, where comparisons can be made, it is found that the agreement between simulations and observations at $z = 1.6$ to $z = 0$ is excellent, as Kolmogorov-Smirnov tests show. It is seen that the peak of the DLA metallicity distribution evolves from $[Z/H] = -1.5$ at $z = 3$–4, to $[Z/H] = -0.75$ at $z = 1.6$, and to $[Z/H] = -0.5$ at $z = 0$. In the used convention, $[Z/H] = 0$ corresponds to solar metallicity value and $[Z/H] = -1$ means one-tenth of solar metallicity, $[Z/H] = -2$ means one hundredth of solar metallicity, etc. Both simulations and nature indicate that there is a weak but real evolution in DLA metallicity. What is also important to note is that, in agreement with observations, simulations indicate that the distribution of metallicity is very wide, spanning three or more decades at $z \geq 1.6$–4. This wide range reflects the rich variety of neutral gas that composes the DLA population, from pristine gas clouds falling onto or feeding galaxies to cold neutral gas clouds in galactic disks. There is a metallicity floor at $[Z/H] \sim -3$ at $z = 1.6$–4, and that floor moves up to $[Z/H] \sim -1.5$ by $z = 0$, consistent with observations (Prochaska et al. 2003). The distribution at $z = 0$ is significantly narrower, partly reflecting the overall enrichment of the IGM and partly due to much reduced variety of DLAs with galactic disks becoming a more dominant contributor to DLAs.

➲ *Figure 12-19* shows the size (radius) distribution at redshift $z = 0, 1.6, 3.1, 4.0$ for individual DLA size and total DLA size of each galaxy. ➲ *Figure 12-20* presents the size information in a different way, where the probability that, for a random pair of sightlines separated by $(30, 20, 10, 5, 3)$ kpc at the redshift in question, one sightline intercepts a DLA and the other intercepts a column density lower than shown in the x-axis. On average, individual DLA size as well as the total DLA size of galaxies are larger at high redshift than at lower redshift. This is consistent with the gradual dominance of contribution from stellar disk with decreasing redshift. The individual DLA size distribution appears to sharply peak at $r_{\mathrm{DLA}} \sim 15$ kpc at $z = 4$, then move left to a broader peak at $r_{\mathrm{DLA}} \sim 10$–12 kpc at $z = 3.1$, then sharpen somewhat to peak at $r_{\mathrm{DLA}} \sim 4$–5 kpc at $z = 1.6$, and finally move rightward slightly to peak at $r_{\mathrm{DLA}} \sim 6$–7 kpc at $z = 0$. What is quite remarkable is that, at $z = 1.6$ and $z \sim 3$ where comparisons can be made, the predicted size distributions and the available observations agree very well. This is a testament to the success of the standard cold dark matter model.

Finally, the model also successfully reproduces the following properties of DLAs: column density distribution evolution, line density evolution, kinematic structural parameters evolution, neutral mass content evolution, and others. Taking all together, it is concluded that the standard cold dark matter model can now satisfactorily explain all observed properties of DLAs. The next frontier of research in this subject area will be to understand what role DLAs play in fueling galaxy formation.

4.5 Metal-Line Systems

One of the pillars of the Big Bang theory is its successful prediction of a primordial baryonic matter composition for the IGM, made up of nearly 100% hydrogen and helium with a trace amount of a few other light elements (e.g., Schramm and Turner 1998). Galaxies, collectively through stellar winds and supernova explosions, return a significant amount of material that formed into stars back to the IGM, including "metals" – in the astronomers' term, all elements in the periodic table except H and He that are nucleosynthesized in stars. Metals are found almost everywhere in the observable IGM, ranging from the metal-rich intracluster medium (e.g., Mushotzky and Loewenstein 1997) to moderately enriched damped Lyman systems (e.g., Pettini et al. 1997; Prochaska et al. 2003) to low metallicity Lyman alpha clouds (e.g., Schaye et al. 2003). It is generally believed that metals in the IGM originate in galaxies. To see how metals produced in galaxies get transported to the IGM via galactic winds powered collectively by stellar winds and supernova explosions, ➲ *Fig. 12-21* shows the spatial distribution of metallicity in the IGM with (left panel) and without (right panel) galactic winds from cosmological hydrodynamic simulations. While other, gravitational and hydrodynamic processes do transport metals to the vicinity ($\leq\sim$ 100 kpc) of galaxies (right panel of ➲ *Fig. 12-5*), galactic winds clearly play a more important role to transport the metals from galaxies to larger distances, in conjunction with other processes. The "metal bubbles" (reddish bubbles seen in the left panel of ➲ *Fig. 12-21*) have the ratio of metal density to total gas density (i.e., metallicity) equal to $\rho_{\mathrm{metals}}/\rho_{\mathrm{gas}} \sim 10^{-4}$, indicating that these metal-contaminated regions are enriched to a metallicity close to about 1% of solar value. Most of the volume far from galaxies, however, remains uncontaminated by galactic winds. What is the metallicity of the IGM in low-density regions that are distant

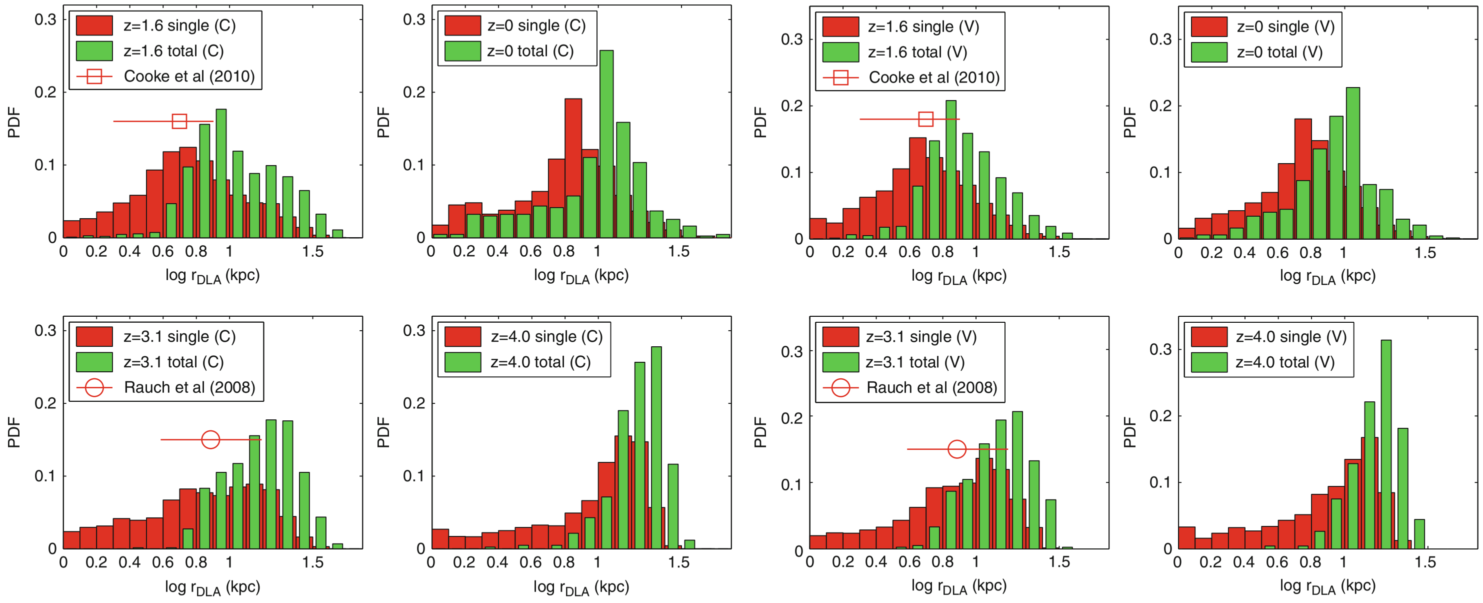


Fig. 12-19

Left set of four panels: the DLA size distribution at redshift $z = 0, 1.6, 3.1, 4.0$ for "C" run. Each individual DLA size r_{DLA} (see text for definition) is shown as *red histograms*, whereas the total DLA size of a galaxy r_{tot} (see text for definition) is shown as *green histograms*. *Right set of four panels*: the DLA size distribution at redshift $z = 0, 1.6, 3.1, 4.0$ for "V" run. The observationally inferred DLA size, shown as an *open square* in both $z = 1.6$ panels, is from Cooke et al. (2010), and that shown as an *open circle* in both $z = 3.1$ panels is from Rauch et al. (2008) with the shown dispersion estimated by this author (This figure is taken from Cen (2012))

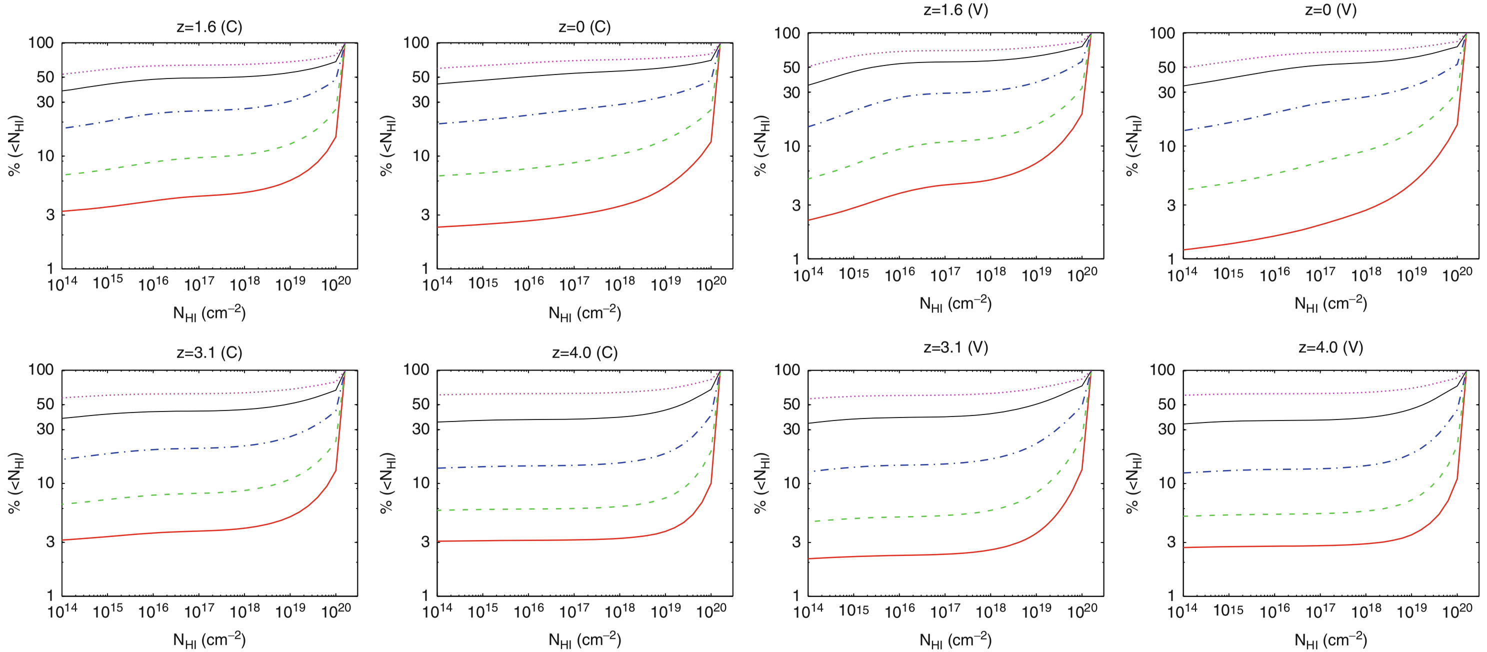


Fig. 12-20

The probability of the second line of sight having an HI column lower than the value shown in the x-axis, while the first line of sight is known to have intercepted a DLA at projected separation of (30, 20, 10, 5, 3) kpc (five curves from top to bottom shown in each panel) at the redshift in question. The *left set* of four panels are at redshift $z = 0, 1.6, 3.1, 4.0$ for "C" run; the *right set* of four panels are for "V" run (This figure is taken from Cen (2012))

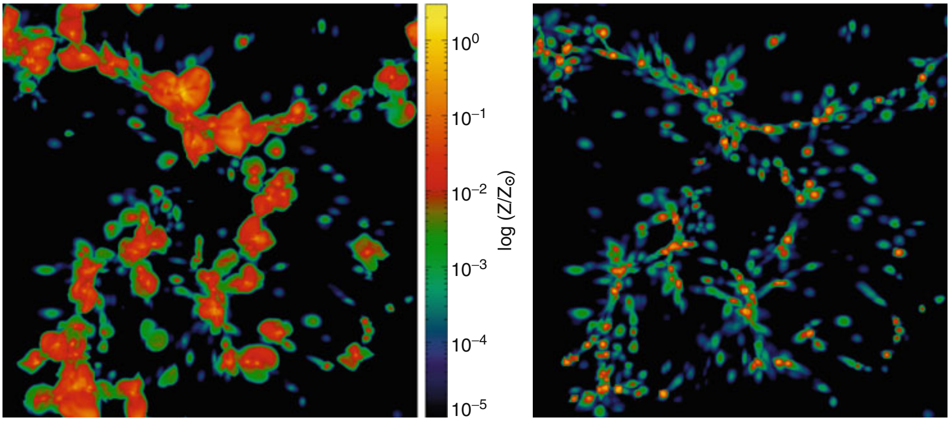


Fig. 12-21

Projected metallicity of a slice of size 11 × 11 h^{-2} Mpc^2 comoving and a depth of 2.75 h^{-1} Mpc comoving at redshift $z = 3$ for a *WMAP*-normalized ΛCDM model with (*left panel*) and without (*right*) galactic winds, respectively. The strength of the galactic winds is normalized to LBG observations (This figure is taken from Cen et al. (2005))

from galaxies? Is there gas that is still primordial? These questions remain unanswered observationally. Theoretically, large uncertainties exist, primarily due to lack of knowledge of galaxy formation at high redshift.

Needless to say, matter that ends up in stars comes from the IGM. In other words, the interactions between galaxies and the IGM are two-way. Thus, probing the IGM, especially with metals, yields vital information about the physical process of galaxy formation and feedback. Astronomers have done this for decades. Metal-line absorption systems in QSO spectra are the primary probes of the metal enrichment of the IGM as well as in the vicinities of galaxies. The most widely used metal lines include Mg II $\lambda\lambda$2796, 2803 doublet; C IV $\lambda\lambda$1548, 1550 doublet; and O VI $\lambda\lambda$1032, 1038 doublet. A brief up-to-date account of the C IV and O VI absorption lines and the global evolution of metals in the IGM in the redshift range $z = 0$–6 is given here in the context of theory confronted with observations. The C IV $\lambda\lambda$1548, 1550 doublet places itself at the redward of Lyα line and hence is easily distinguished from the Lyα forest lines. The ionization potential of O VI and the relatively high oxygen abundance are very favorable for production of O VI absorbers in the IGM. The rest wavelength of OVI (1,032, 1,037 Å) places it within the Ly-α forest, which makes the identifications of these lines more complicated. O VI absorption lines can probe the metal content of the IGM in ways complementary to what is provided by C IV lines. For example, the O VI lines can probe IGM that is hotter than that probed by the C IV lines and can reach lower densities thanks to higher abundance.

For a statistically understood sample of QSO absorption lines, one could derive the cosmological density contained in them. Early investigations indicate that $\Omega_{\rm CIV}$ remains approximately constant in the redshift interval $z \sim 1.5$–4 (e.g., Songaila 2005). There have been recent efforts to extend the measurements of $\Omega_{\rm CIV}$ to $z < 1.5$ (Cooksey et al. 2010) and to $z > 5$ (Simcoe 2006; Ryan-Weber et al. 2006, 2009; D'Odorico et al. 2010; Becker et al. 2009). Observations in these redshift ranges have been difficult to carry out because C IV transition moves to the UV at low redshift and to the IR band at high redshift. D'Odorico et al. (2010) find evidence of a rise in the C IV mass density for $z < 2.5$. Simcoe (2006) and Ryan-Weber et al. (2006) found evidence of C IV density at $z \sim 6$ being consistent with estimations at $z \sim 2$–4.5. More recently, however, Becker et al. (2009) set upper limits for Ω_{CIV} at $z \sim 5.3$ and Ryan-Weber et al. (2009) observe a decline in intergalactic C IV approaching $z = 6$, which are in fact in good agreement with latest cosmological hydrodynamic simulations, as will be shown below. More recently, observational studies of O VI absorbers and measurements of the O VI mass density, Ω_{OVI}, have been made (Carswell et al. 2002; Bergeron et al. 2002; Simcoe et al. 2004; Simcoe 2006; Frank et al. 2012).

Like for Lyα clouds, individual metal absorption lines and their properties may be characterized. *Figure 12-22* shows the column density distributions for C IV and O VI absorbers at $z = 2.5$. Overall, the predictions of the standard cold dark matter model are in good agreement with extant observations for both C IV and O VI lines over the range of column density shown, $N = 10^{13} - 10^{15}\ \mathrm{cm}^{-2}$, given the uncertainties of the current observational data. Since the regions probed by C IV lines and O VI lines are often physically different and to some extent reflect the different stages of the evolution of the feedback shocks, the good agreement between the simulations and observations suggests that current physical treatment of the feedback process provides a good approximation to nature, and it is indirect but strong evidence that feedback from star formation plays the central role in enriching the IGM with its energy and metals.

Metals in the IGM ultimately comes from stars. It is tempting to extract information from the evolution of metal density in the IGM to infer the star formation history. *Figure 12-23* shows the evolution of the mass density contained in the C IV (left) and O VI (right) absorption lines, respectively. Considering the observational uncertainties and cosmic variance, it is very encouraging to see the good agreement between the simulation results and observations over the entire redshift range $z \sim 2$–6, where comparisons may be made. In agreement with observations, the mass density contained in the C IV absorption line is, within a factor of 2, constant

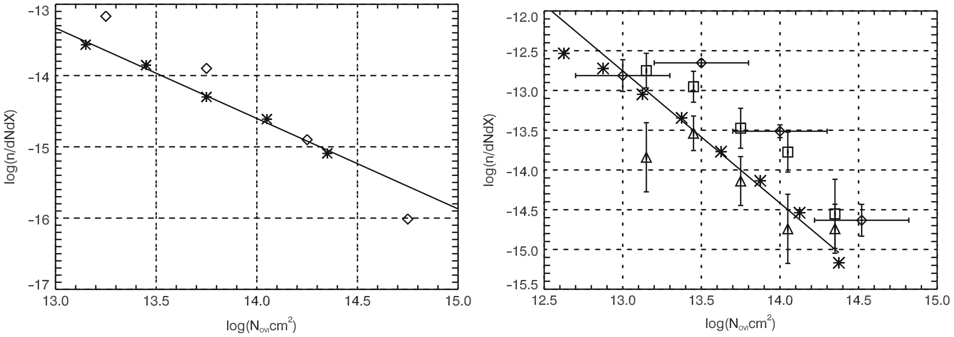


Fig. 12-22

***Left panel*: the computed column density distribution for the C IV absorption line at $z = 2.5$ is shown as the *solid line*, which is the best power-law fit to our simulated results shown in asterisks. The slope of the fit is −1.269 ± 0.061. Diamonds are observational data from Songaila (2005) corrected for our cosmology. *Right panel*: the computed column density distribution for the O VI absorption line at $z = 2.5$ is shown as the *solid line*, which is the best power-law fit to our simulated results, with slope −1.662 ± 0.093. The observational data are from Carswell et al. (2002) (*squares*), Bergeron and Herbert-Fort (2005) (*diamonds*), and Simcoe et al. (2002) (*triangles*) corrected for our cosmology (This figure is taken from Cen and Chisari (2011))**

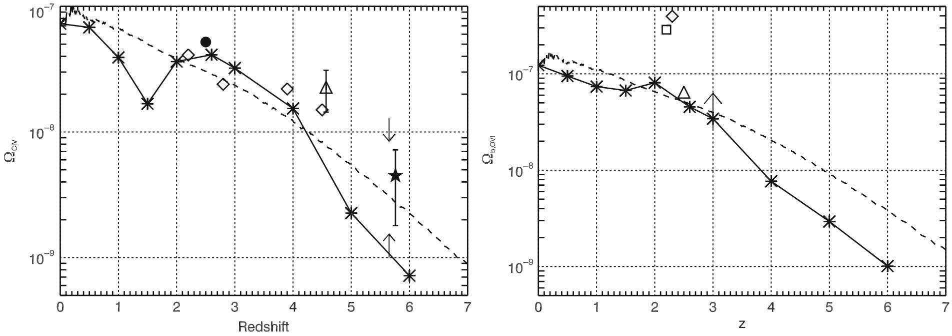


Fig. 12-23

***Left panel*: redshift evolution of Ω_{CIV} from simulation (*asterisks*). Observational data are from Songaila (2005) (*open diamonds*), Becker et al. (2009) (*arrows as limits*), Pettini et al. (2003) (*open triangle*), Ryan-Weber et al. (2009) (*filled star*), and Simcoe (2006) (*filled circle*). The *dashed curve* is a simple physical model to explain the evolution of Ω_{CIV} (see Cen and Chisari 2011). *Right panel*: redshift evolution of Ω_{OVI}. Observational data are from Carswell et al. (2002) (*open square*), Bergeron and Herbert-Fort (2005) (*open diamond*), Simcoe et al. (2002) (*open triangle*) and Frank et al. (2012) (*lower limit, arrow*) (This figure is taken from Cen and Chisari (2011))**

from $z = 1$ to $z = 4$ (e.g., Songaila 2001, 2005) and subsequently drops by a factor of ~$(10, 40)$ by $z = (5, 6)$ (Becker et al. 2009; Ryan-Weber et al. 2009). Because the total amount of star formation and metals in the IGM is observed to have increased significantly in the redshift range $z = 0$–4, it seems that the near constancy of Ω_{CIV} at the redshift range $z = 1$–4 and Ω_{OVI} at the redshift range $z = 0$–2 reflects neither the star formation history nor the amount of metals in the IGM. This reflects a "selection effect" of said absorption systems of the overall metals in the IGM, which may be due to a combination of several different processes, including the evolution of the mean gas density as $(1+z)^3$, the evolution of the overdensity of the regions that produce C IV lines, the density dependence of the IGM metallicity and its evolution, the evolution of the radiation background, and hierarchical build-up hence gravitational shock heating of the large-scale structure. Both the magnitude and evolution of Ω_{OVI} are somewhat different from C IV absorbers.

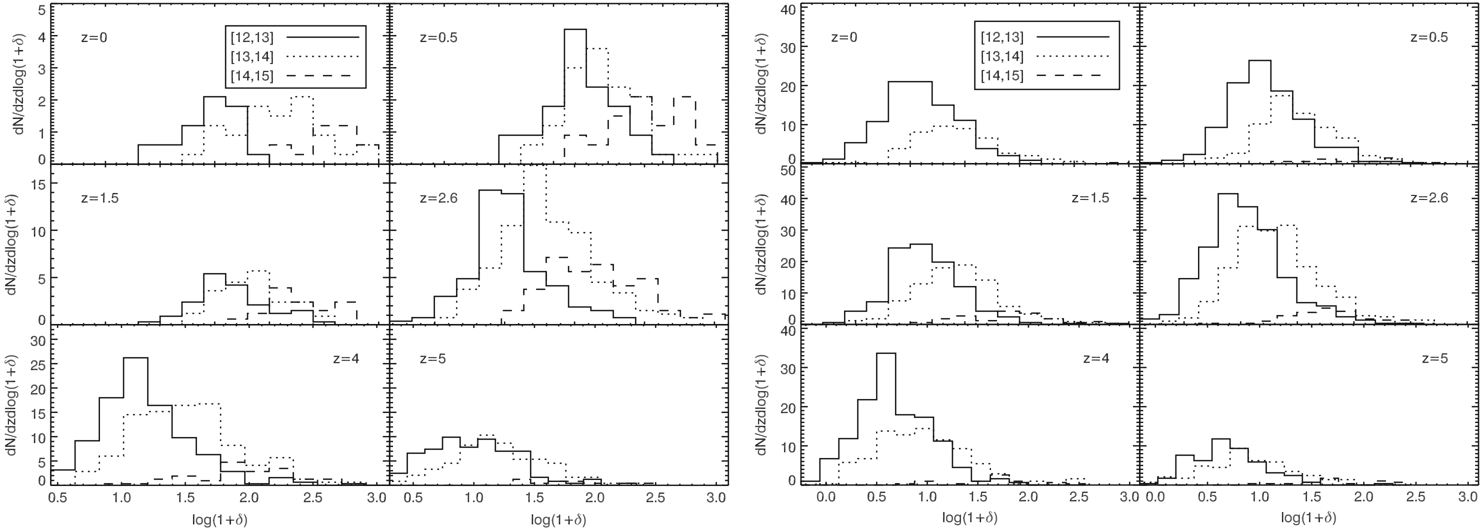
To understand the physical properties of C IV and O VI absorbers, different projections through the multidimensional parameter space spanned by several fundamental physical variables are now shown. *Figure 12-24* shows the distribution of gas overdensity for C IV (left) and O VI absorbers (right) at six different redshifts, $z = (0, 0.5, 1.5, 2.6, 4, 5)$. A comparison of the three histograms for three subsets of C IV and O VI absorbers in each panel indicates that higher column C IV and O VI absorbers are produced, on average, by higher density gas. There is a clear trend that C IV absorbers trace increasingly more overdense regions with decreasing redshift. For example, while the location of the vast majority of C IV absorbers with $\log(N_{\mathrm{C\,IV}}\ \mathrm{cm}^2) = [12, 13]$ appears to be outside virialized regions (i.e., overdensity less than about 100) at $z > 2.6$, a significant fraction of them reside in virialized regions at $z < 1.5$; the same is true for higher column C IV absorbers. A comparison to O VI absorbers reveals a striking contrast: the vast majority of O VI absorbers with $\log(N_{\mathrm{C\,IV}}\ \mathrm{cm}^2) \le 14$ are located outside virialized regions *at all redshifts*. In addition, typical O VI lines arise from somewhat lower density regions than C IV lines. For example, for O VI absorbers of $\log(N_{\mathrm{C\ IV}}\ \mathrm{cm}^2) = [12, 13]$, the typical overdensity peaks at $\delta \sim 5$ for O VI absorbers versus ~10 for C IV lines at $z = 2.6$–5, which jumps to $\delta \sim 10$ for O VI absorbers versus ~50 for C IV absorbers at $z = 1.5$.

➲ *Figure 12-25* shows the distribution of gas metallicity for C IV (left) and O VI absorbers (right) at six different redshifts, $z = (0, 0.5, 1.5, 2.6, 4, 5)$. It is seen that C IV absorption lines arise from gas with a wide range of metallicity from [C/H] = −3 to −0.5, peaked approximately around −2.5 to −1.5 at $z > 0.5$. The metallicity distributions for O VI absorbers are generally cut off at a higher metallicity than those for C IV absorbers at the low end and peak at a higher metallicity. The situation appears to start reversing at $z = 1.5$ such that at $z < 0.5$ the fraction of high metallicity C IV absorbers exceeds that of O VI absorbers. What is also interesting is that the typical metallicity of C IV and O VI lines displays a nonmonotonic trend at a fixed column density. For O VI absorbers, at $z = 4$–5 the metallicity of O VI lines with $\log(N_{\mathrm{O\ VI}}\ \mathrm{cm}^2) = [12, 14]$ peaks at $[Z/Z_\odot] = -1.5$ to -1.0, which moves to a lower value of $[Z/Z_\odot] = -2.0$ to -1.5 at $z = 2.6$, then slightly moves back up to $[Z/Z_\odot] \sim -1.5$ at $z = (1.5, 0.5, 0)$. For comparison, the overall behavior for C IV lines is as follows: the metallicity of C IV lines with $\log(N_{\mathrm{O\ VI}}\ \mathrm{cm}^2) = [12, 14]$ peaks at $Z = -2.0$ to -1.5 at $z = 5$, at $[Z/Z_\odot] \sim -2$ at $z = 4$, followed by a very broad distribution peaking at $Z = -2$ to -1 at $z = 1.5$ to $z = 2.6$ with a larger fraction reaching a relatively high metallicity gas with $[Z/Z_\odot] > -1$.

The overall trend in metallicity evolution with redshift for the C IV and O VI absorbers could be understood as follows. First, note that the ionizing radiation background at $z = 4, 5$ is about (1/3, 1/30) of that $z = 2.6$, which in turn is larger than that at $z = (1.5, 0.5, 0)$ by a factor of $\sim(2, 7, 30)$. At $z = 4$–5, both C IV and O VI absorbers are predominantly collisionally ionized with the temperatures peaking at 10^5 and $10^{5.5}$ K, respectively, as shown below in ➲ *Fig. 12-26*. These regions are relatively closer to galaxies, from which metal-carrying shocks originate and have relatively high metallicities. At lower redshift $z = 2.6$, larger regions around galaxies have been enriched with metals and the rise of the ionizing radiation background produces a large population of photoionized C IV and O VI lines at lower temperature and lower metallicity. Toward still lower redshift $z = 1.5$, the decrease of the mean gas density in the universe demands a rise in overdensity of the O VI -bearing gas in order to produce a comparable column density, causing a shift of these regions to be closer to galaxies where both metallicity and density are higher.

The combination of lower density (➲ *Fig. 12-24*) and higher metallicity (➲ *Fig. 12-25*) for the typical (low) column density O VI absorbers compared to C IV absorbers is reminiscent of metal-carrying shocks propagating through inhomogeneous medium, exactly the situation one would expect of the feedback shocks from galaxies entering the highly inhomogeneous IGM. Given the widespread steep density gradients (steeper than −2) in regions just outside the virial radius of galaxies, these shocks could not only heat up lower density regions to higher temperatures but also enrich them to higher metallicity. The feedback shocks generically propagate in a direction that has the least resistance and is roughly perpendicular to the orientation of a local filament where a galaxy sits, as seen clearly in ➲ *Fig. 12-21*. While higher density regions, on average, tend to have higher metallicity, the dispersion is sufficiently large that the reverse and other complex situations often occur in some local regions. This appears to be what is happening here, at least for some regions that manifest in C IV and O VI lines.

➲ *Figure 12-26* shows the distribution of gas temperature for C IV (left) and O VI (right) absorbers. The temperatures of C IV absorbers at $z = 5$ and O VI absorbers at $z = 4$–5 narrowly peak at 10^5 and $10^{5.5}$ K, respectively, suggesting that collisional ionization makes the dominant



■ **Fig. 12-24**

Left panel **shows the distribution of gas overdensity of regions that produce the CVI absorption lines at six different redshifts, $z = 0, 0.5, 1.5, 2.6, 4, 5$, separately for three subsets of lines of column density in the range of $\log N_{\mathrm{C\,IV}}$ cm^2=[12,13],[13,14],[14,15], respectively.** ***Right panel*** **shows the counterpart for O VI absorption lines (This figure is taken from Cen and Chisari (2011))**

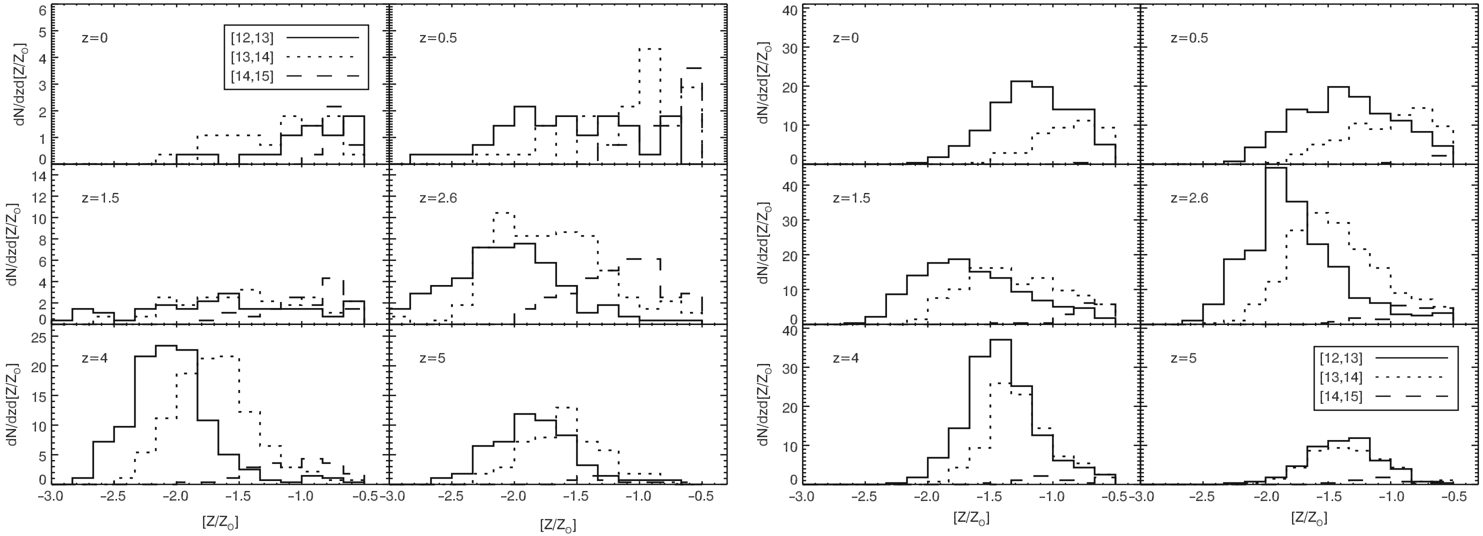
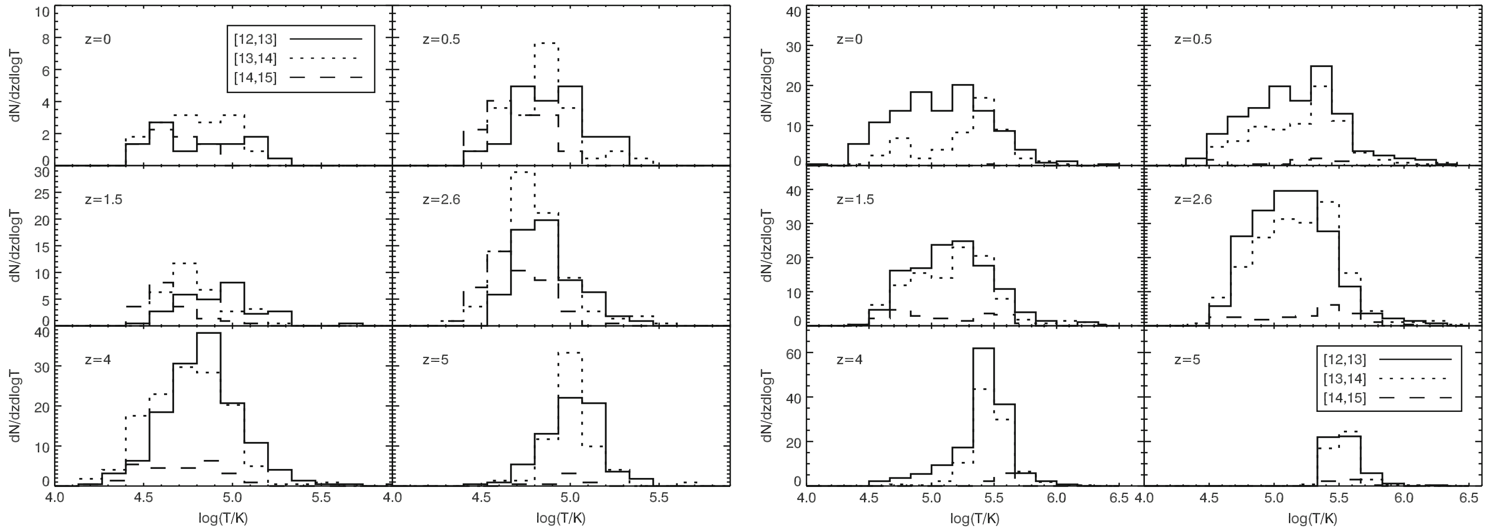


Fig. 12-25

Left panel **shows the distribution of gas metallicity in solar units of regions that produce the CVI absorption lines at six different redshifts, *z* = 0, 0.5, 1.5, 2.6, 4, 5, separately for three subsets of lines of column density in the range of log$N_{\mathrm{C\ IV}}$ cm^2=[12,13],[13,14],[14,15], respectively. *Right panel* shows the counterpart for O VI absorption lines (This figure is taken from Cen and Chisari (2011))**



■ **Fig. 12-26**

Left panel **shows the distribution of gas temperature of regions that produce the C IV absorption lines at six different redshifts, $z = 0, 0.5, 1.5, 2.6, 4, 5$, separately for three subsets of lines of column density in the range of $\log N_{\text{C IV}}$ cm^2=[12,13],[13,14],[14,15], respectively.** ***Right panel*** **shows the counterpart for O VI absorption lines (This figure is taken from Cen and Chisari (2011))**

contribution to both species and the two types of absorbers arise from different regions. The rapid drop in the amplitude of the UV radiation background beyond $z = 3$ and increase in gas density with $(z+1)^3$ are the primary reasons for diminished component of photoionized C IV and O VI absorbers at these high redshifts. At redshift $z < 2.6$, the distributions for the two absorbers become progressively broader ranging from $10^{4.3}$ to $10^{5.5}$ K for C IV absorbers, and from $10^{4.3}$ to 10^6 K for O VI absorbers. Thus, at $z < 2.6$, both C IV and O VI absorbers are a mixture of photoionized and collisionally ionized ones. For both C IV and O VI lines, while the temperature distributions of O VI lines at $z < 2.6$ are broad, there is no significant segregation in temperature of lines of different column densities. Recall that there is a noticeable correlation between column density and overdensity for both O VI lines and C IV lines. This is indicative of complex, inhomogeneous nature of metal-enrichment process around galaxies.

4.6 Warm-Hot Intergalactic Medium

The Warm-hot intergalactic medium is generally defined to be gas in the temperature range of 10^5–10^7 K and is a cosmologically and astrophysically important component of the IGM. Here, the attention is focused on WHIM at $z = 2$–3 and $z = 0$.

4.6.1 The Missing Metals Problem at $z = 2$–3

At $z \sim 2$–3, there is a so-called "*missing metals problem*": observationally, integrating the observed star formation rate history from high redshift down to $z = 2.5$ suggests that the vast majority (possibly ≥80%) of cosmic metals at $z \sim 2.5$ appear to be missing (e.g., Pettini 1999). Note that this conclusion is insensitive to the choice of the stellar initial mass function (IMF), since both UV light and metals are, in the zeroth order, produced by the same massive stars. Metals that have been accounted for in the estimates include those in stars of Lyman break galaxies (LBG), damped Lyman alpha systems (DLAs), and Lyα forest, i.e., cold-warm gas and stars. Thus, all the regions where metals have been accounted for trace temperatures lower than about 3×10^4 K. Since metals are transported to large distances from galaxies by galactic winds as shown in ➋ *Fig. 12-21*, one may ask, could a significant fraction of metals that accompanies the galactic winds energy be heated up and be in a phase like WHIM that is different from those where metals have been inventoried? This is in fact borne out in simulations. ➋ *Figure 12-27* displays the temperature distributions corresponding to the same two projections shown in ➋ *Fig. 12-21*. It is clear that the metal bubbles are generally heated to temperatures higher than about 3×10^4 K, mostly falling in the temperature range of WHIM. Put it simply, it is in WHIM where most of the energy and metal exchanges between galaxies and the IGM take place. It has been proposed earlier that the missing metals may be found in hot gaseous halos of star-forming galaxies (Pettini 1999; Ferrara et al. 2005). That proposal was partly correct, except that simulations show that most of the missing metals at $z = 2$–3 are in a diffuse IGM of temperature in the range $T = 10^{4.5-7}$ K that are *outside of dark matter halos*.

To be more quantitative, the cold-warm IGM component (1) ($T < 10^5$ K cold-warm gas) shown in ➋ *Fig. 12-7* is broken up into two sub-components with (1C) ($T < 3 \times 10^4$ K cold gas) and (1W) ($T = 3 \times 10^4$–10^5 K warm gas). The purpose of this finer division is to separate out

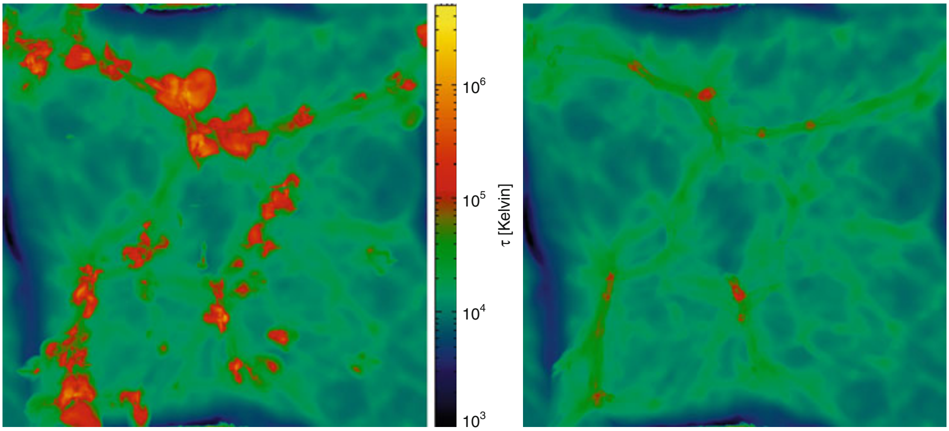


Fig. 12-27

Projected temperature of a slice of size 11 × 11 h^{-2} Mpc^2 comoving and a depth of 2.75 h^{-1} Mpc comoving at redshift z = 3 with (*left panel*) and without (*right*) GSW, respectively. The strength of the GSW is normalized to LBG observations (This figure is taken from Cen et al. (2005))

the cold gas (1C), which can be more appropriately identified with Lyα forest clouds and DLAs. The results are shown in *Fig. 12-28*. About one-third of all metals produced by $z = 0$ is locked up in stars, decreasing monotonically toward high redshift, dropping to about 10% by $z = 5$. The fraction of metals in the hot X-ray emitting component is at about 10% level at $z = 0$, plummeting to about 2% at $z = 2$ and slowly rising back to about 6% at $z = 6$. The remaining metals are in the general photoionized Lyα forest and the WHIM. At $z = 6$, the Lyα forest ($T < 3 \times 10^4$ K, open triangles) contains about 43% of all metals, while WHIM ($T = 10^5$–10^7 K, open circles) and warm IGM ($T = 3 \times 10^4$–10^5 K, solid triangles) contain 39% and 7%, respectively. But the fraction of metals in the Lyα forest decreases steadily with time and becomes a minor component by $z = 0$ at $< 3\%$. Most of the metals are seen to be contained in the WHIM at all times below redshift five at 50–60%, peaking at ~60% at redshift $z \sim 2$. In total, the amount of metals contained in the IGM with temperature $T > 3 \times 10^4$ constitutes about 2/3 of all metals produced by $z = 2.5$. Metals in this temperature range have not been accounted for in the quoted observational inventory at $z = 2.5$.

What is the typical density of this WHIM gas where most of the metals are hidden? The distribution of metal mass as a function of density for each IGM component is shown in *Fig. 12-29*. A very interesting result is that at high redshift ($z = 3, 5$) the metal mass in the WHIM tends to peak at a somewhat lower overdensity than that for the overall WHIM mass, thanks to the upturn of metallicity of the WHIM at low overdensity end. Specifically, at $z = 3 - 5$, it appears that the metal mass peaks at $\delta \sim 2$, whereas the total WHIM mass peaks at $\delta \sim 10$. This trend is reversed at lower redshift; for example, at $z = 0$, the metals in WHIM is now broadly peaked at $\delta \sim 100$, while the WHIM mass peaks at $\delta \sim 10$. This reversal is due to accretion of metal-enriched gas onto high-density regions during recent formation of large-scale structures. Quantitatively, at $z = 2.5$, only about 15% of the metals in warm and WHIM gas is located within virialized regions. About 73% of the metals in warm and WHIM gas resides

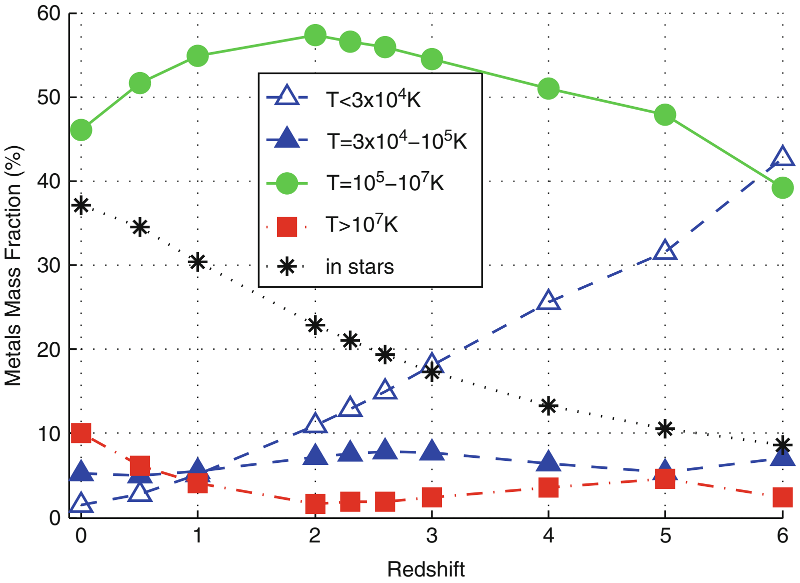


Fig. 12-28
Shows the evolution of fractions of all metals produced that are contained in each of the five components, as a function of redshift: (1C) $T < 3 \times 10^4$ K cold gas, (1W) $T = 3 \times 10^4$–10^5 K warm gas, (2) WHIM at 10^7 K > $T > 10^5$ K, (3) Hot X-ray emitting gas at $T > 10^7$ K, (4) "stars" (This figure is taken from Cen and Chisari (2011))

in the IGM with $\delta = 1$–100, with the remaining 12% in underdense regions. Thus, the solution to the long-standing problem of missing metals at $z = 2$–3 has been found: *most of the missing metals are in the warm and WHIM gas with moderate overdensity broadly distributed between* $\delta \sim 1$–10.

4.6.2 The Missing Baryons Problem at $z = 0$

In the remarkably successful standard cosmological model (Krauss and Turner 1995), most of the mass-energy density in the universe is in dark energy and dark matter with cold dark matter being the most popular choice for the latter. But the puzzle extends beyond our ignorance of the nature of the dynamically measured matter. The remainder, 4–5%, is in the normal matter – baryons, as detected by the WMAP cosmic microwave background (CMB) observations (Komatsu et al. 2011) at $z \sim 1{,}100$ and verified at much lower redshift ($z \sim 3$) by the hydrogen Lyα absorption observations (Weinberg et al. 1997; Rauch et al. 1997):

$$\Omega_{\text{baryon}}(\text{Ly}\alpha) \geq 0.017\,\text{h}^{-2} = 0.035. \tag{12.2}$$

In addition, the observed light-element ratios (in particular, the deuterium to hydrogen ratio) in some carefully selected absorption line systems at $z = 2$–3, interpreted within the context of the standard light element nucleosynthesis theory, yield the total baryonic density

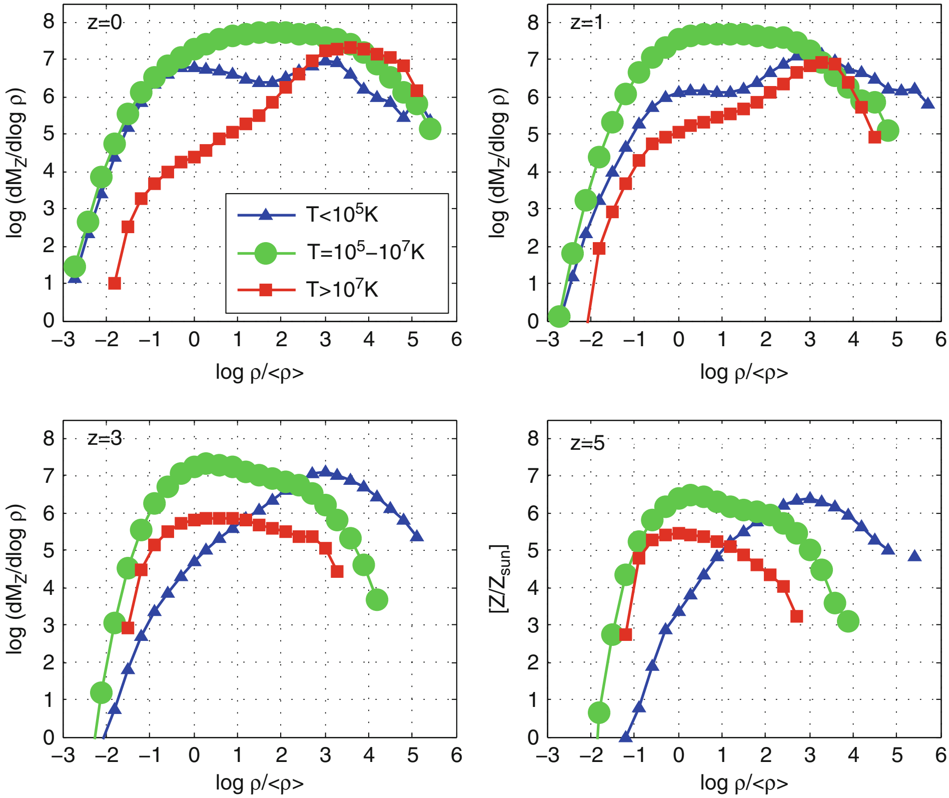


Fig. 12-29

Shows the distributions of metal mass for the three IGM components – (1) cold-warm gas at $T < 10^5$ K, (2) WHIM at 10^7 K > $T > 10^5$ K, and (3) Hot X-ray emitting gas at $T > 10^7$ K – as a function of overdensity at four different redshifts $z = 0, 1, 3, 5$. Note that the area under each curve is proportional to the metal mass contained (This figure is taken from Cen and Chisari (2011))

(Burles and Tytler 1998; Kirkman et al. 2003; Pettini et al. 2008) that is consistent with those inferred by CMB and Lyα forest observations:

$$\Omega_{\text{baryon}}(D/H) = (0.019 \pm 0.001)\,\text{h}^{-2} = 0.039 \pm 0.002. \tag{12.3}$$

The agreement between these three completely independent measurements is remarkable. But, at redshift zero, after summing over all well-observed contributions, the baryonic density appears to be far (by a factor greater than 3) below (e.g., Fukugita et al. 1998) this level:

$$\Omega_{\text{b}}(z = 0)|\text{seen} = \Omega_{*} + \Omega_{HI} + \Omega_{H_2} + \Omega_{\text{Xray,cl}} \approx 0.0068 \leq 0.011 \;\; (2\sigma \text{ limit}), \tag{12.4}$$

where Ω_{*}, Ω_{HI}, Ω_{H_2}, and $\Omega_{\text{Xray,cl}}$ are the baryonic densities contained in stars, neutral atomic hydrogen, molecular hydrogen, and hot X-ray emitting gas in rich cluster centers, respectively, in units of the critical density. Thus, unless three independent errors have been made in the arguments that led to the three measurements, there is a sharp decline of the amount of observed

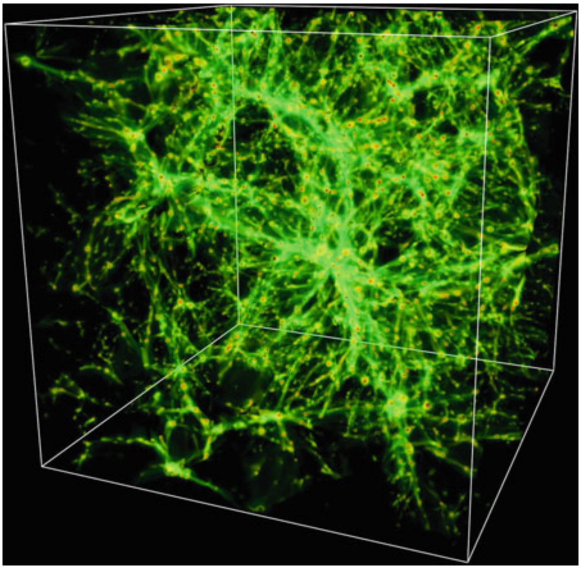


Fig. 12-30

Spatial distribution of the warm/hot gas with temperature in the range 10^5–10^7 K at $z = 0$. The *green regions* have densities about ten times the mean baryon density of the universe at $z = 0$; the *yellow regions* have densities about 100 times the mean baryon density, while the small isolated regions with red and *saturated dark colors* have even higher densities reaching about 1,000 times the mean baryon density and are sites for current galaxy formation (This figure is taken from Cen and Ostriker (1999))

baryons from high redshift to the present-day; i.e., most of the baryons in the present-day universe are missing. A significant fraction of missing baryons locally may have been found in the Lyα forest primarily in warm gas with $\Omega_{\mathrm{Ly}\alpha,z=0} \sim 0.013$ (e.g., Penton et al. 2004; Danforth and Shull 2008). But about 50% of all baryons are still missing. A simple prediction from cosmological hydrodynamic simulations (e.g., Cen and Ostriker 1999, 2006; Davé et al. 2001) is that nearly one-half of all baryons in the local universe should be in cosmic web of WHIM with densities of 10–300 times the mean baryon density, as shown in ➋ *Fig. 12-30*.

The O VI (1,032, 1,034) Å doublet line provides vital information on the WHIM. The reality of the WHIM, at least the low-temperature portion of it ($T \leq 10^6$ K), has now been convincingly confirmed by a number of observations from HST and FUSE, through the O VI absorption line in the FUV portion of QSO spectra (e.g., Tripp et al. 2000; Danforth and Shull 2005). The overall agreement between theory and observations with respect to O VI absorption line is excellent, as typified by the column density distribution shown in ➋ *Fig. 12-31*. The part of WHIM detected in O VI absorption constitutes about 20% of total WHIM. The Cosmic Origins Spectrograph (COS) now aboard the Hubble Space Telescope (HST) has a sensitivity 10–30 times greater than previous UV spectrographs on HST. COS will provide at least an order of magnitude increase in the number of detected O VI absorption lines (Shull 2009) to provide unprecedented statistical power to further scrutinize theory to test both the cosmological model as well as galaxy formation.

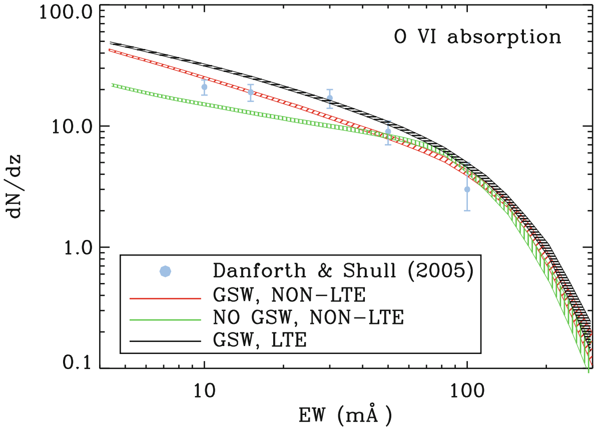


Fig. 12-31
Shows the number of O VI absorption lines per unit redshift as a function of equivalent width in units of mÅ. The *red curve* shows our primary results from the simulation with GSW and the *green curve* from the simulation without GSW. The *black curve* is computed using CLOUDY code on the assumption of ionization equilibrium based on the density, temperature, and metallicity information from the simulation with GSW. The symbols are observations by Danforth and Shull (2005) (This figure is taken from Cen and Fang (2006))

The search for X-ray absorption of WHIM, in the higher temperature portion ($T \geq 10^6$ K), turns out to be elusive. The O VII absorber at $\lambda = 21$ Å is predicted to be the most abundant and contains a large portion of the WHIM gas (e.g., Hellsten et al. 1998; Cen and Fang 2006), and it is therefore extremely important to be able to detect O VII WHIM gas to account for the missing baryons. Early pioneering observations (e.g., Fang et al. 2001) produced no convincing detections. Mathur et al. (2003) performed a dedicated deep observation of the quasar H 1821+643, which has several confirmed intervening O V I absorbers, but found no significant X-ray absorption lines at the redshifts of the O V I systems. Nicastro et al. (2005) embarked on a campaign to observe Mrk 421 during its periodic X-ray outbursts with the Chandra LETGS and presented evidence for the detection of two intervening absorption systems at $z = 0.011$ and $z = 0.027$. But the spectrum of the same source observed with the XMM-Newton RGS does not show these absorption lines (Rasmussen et al. 2007), despite higher signal-to-noise and comparable spectral resolution. Kaastra et al. (2006) have reanalysed the Chandra LETGS data and are in agreement with Rasmussen et al. (2007). Observations of 1ES 1028+511 at $z = 0.361$ by Steenbrugge et al. (2006) yield no convincing evidence for WHIM absorption. Theoretical predictions also suggest that high-temperature portion of the WHIM should be closely associated with groups and clusters. So far, observations performed to detect the absorption by WHIM associated with known massive clusters are indeed more successful. An XMM-Newton RGS spectrum of quasar LBQS 1228+1116 revealed a marginal feature at the Virgo redshifted position of O VIII Lyα at the 95% confidence level (Fujimoto et al. 2004). Using XMM-Newton RGS observations of an active galactic nuclei behind the Coma Cluster, X Comae, Takei et al. (2007) claimed to have detected WHIM associated with the Coma cluster. Through the Sculptor Wall, Buote et al. (2009) have detected WHIM O VII absorption at a column greater than 10^{16} cm^{-2}. To fully settle the issue of the high-temperature end ($T \geq 10^6$ K) of WHIM, a high-throughput and high-sensitivity X-ray mission for absorption study will be needed. Direct imaging of the emission lines in X-ray, such as the O VII 21 Å line, will be extremely valuable (e.g., Ohashi et al. 2010).

In summary, the WHIM, as a conduit for exchanges of matter and energy between galaxies and the general IGM, plays a very important role in galaxy formation. Ongoing observational and theoretical efforts in this area are expected to significantly advance our understanding of galaxy formation and its feedback.

Acknowledgments

I would like to thank my collaborators who have made indispensable contributions to many of the works on which this review is substantially based, including Eliza Chisari Taotao Fang, Patrick McDonald, Jordi Miralda-Escude, Kentaro Nagamine, Jeremiah P. Ostriker, Michael Rauch, Hy Trac, and John Wise.

References

Barkana, R., & Loeb, A. 2001, Phys. Rep., 349, 125

Becker, G. D., Rauch, M., & Sargent, W. L. W. 2009, ApJ, 698, 1010

Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Spergel, D. N., Tucker, G. S., Wollack, E., Wright, E. L., Barnes, C., Greason, M. R., Hill, R. S., Komatsu, E., Nolta, M. R., Odegard, N., Peiris, H. V., Verde, L., & Weiland, J. L. 2003, ApJS, 148, 1

Bergeron, J., Aracil, B., Petitjean, P., & Pichon, C. 2002, A&A, 396, L11

Bergeron, J., & Herbert-Fort, S. 2005, Probing Galaxies through Quasar Absorption Lines, IAU Colloquium Proceedings of the International Astronomical Union 199, held March 14–18, Shanghai, People's Republic of China, ed. Peter R. Williams, Cheng-Gang Shu and Brice Menard. (Cambridge: Cambridge University Press) pp. 265–280

Bromm, V., Ferrara, A., Coppi, P. S., & Larson, R. B. 2001, MNRAS, 328, 969

Buote, D. A., Zappacosta, L., Fang, T., Humphrey, P. J., Gastaldello, F., & Tagliaferri, G. 2009, ApJ, 695, 1351

Burles, S., & Tytler, D. 1998, ApJ, 499, 699

Carswell, B., Schaye, J., & Kim, T. 2002, ApJ, 578, 43

Cen, R. 2003, ApJ, 591, 12

Cen, R. 2012, The Astrophysical Journal, Volume 748, Issue 2, article id. 121

Cen, R., & Chisari, N. E. 2011, ApJ, 731, 11

Cen, R., & Fang, T. 2006, ApJ, 650, 573

Cen, R., Miralda-Escude, J., Ostriker, J. P., & Rauch, M. 1994, ApJ, 437, L9

Cen, R., Nagamine, K., & Ostriker, J. P. 2005, ApJ, 635, 86

Cen, R., & Ostriker, J. P. 1999, ApJ, 514, 1

Cen, R., & Ostriker, J. P. 2006, ApJ, 650, 560

Cen, R., & Simcoe, R. A. 1997, ApJ, 483, 8

Cole, S., Norberg, P., Baugh, C. M., Frenk, C. S., Bland-Hawthorn, J., Bridges, T., Cannon, R., Colless, M., Collins, C., Couch, W., Cross, N., Dalton, G., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Madgwick, D., Peacock, J. A., Peterson, B. A., Sutherland, W., & Taylor, K. 2001, MNRAS, 326, 255

Cooke, R., Pettini, M., Steidel, C. C., King, L. J., Rudie, G. C., & Rakic, O. 2010, Monthly Notices of the Royal Astronomical Society, Volume 409, Issue 2, pp. 679–693

Cooksey, K. L., Thom, C., Prochaska, J. X., & Chen, H. 2010, The Astrophysical Journal, Volume 708, Issue 1, pp. 868–908

Danforth, C. W., & Shull, J. M. 2005, ApJ, 624, 555

Danforth, C. W., & Shull, J. M. 2008, ApJ, 679, 194

Davé, R., Cen, R., Ostriker, J. P., Bryan, G. L., Hernquist, L., Katz, N., Weinberg, D. H., Norman, M. L., & O'Shea, B. 2001, ApJ, 552, 473

D'Odorico, V., Calura, F., Cristiani, S., & Viel, M. 2010, Monthly Notices of the Royal Astronomical Society, Volume 401, Issue 4, pp. 2715–2721

Dunkley, J., Komatsu, E., Nolta, M. R., Spergel, D. N., Larson, D., Hinshaw, G., Page, L., Bennett, C. L., Gold, B., Jarosik, N., Weiland, J. L., Halpern, M., Hill, R. S., Kogut, A., Limon, M., Meyer, S. S., Tucker, G. S., Wollack, E., & Wright, E. L. 2009, ApJS, 180, 306

Fan, X., Strauss, M. A., Becker, R. H., White, R. L., Gunn, J. E., Knapp, G. R., Richards, G. T., Schneider, D. P., Brinkmann, J., & Fukugita, M. 2006, AJ, 132, 117

Fang, T., Marshall, H. L., Bryan, G. L., & Canizares, C. R. 2001, ApJ, 555, 356

Ferrara, A., Scannapieco, E., & Bergeron, J. 2005, ApJ, 634, L37

Frank, S., Mathur, S., & York, D. G. 2012, The Astronomical Journal, Volume 140, Issue 3, pp. 817–834

Fujimoto, R., Takei, Y., Tamura, T., Mitsuda, K., Yamasaki, N. Y., Shibata, R., Ohashi, T., Ota, N., Audley, M. D., Kelley, R. L., & Kilbourne, C. A. 2004, PASJ, 56, L29

Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998, ApJ, 503, 518

Gardner, J., Katz, N., Hernquist, L., & Weinberg, D. H. 1997a, ApJ, 484, 31

Haiman, Z., & Holder, G. P. 2003, ApJ, 595, 1

Hellsten, et al. 1998, The Astrophysical Journal, Volume 509, Issue 1, pp. 56–61

Hernquist, L., Katz, N., Weinberg, D. H., & Jordi, M. 1996, ApJ, 457, L51

Hong, S., Katz, N., Davé, R., Fardal, M., Kereš, D., & Oppenheimer, B. D. 2010, ArXiv e-prints

Kaastra, J. S., Werner, N., Herder, J. W. A. d., Paerels, F. B. S., de Plaa, J., Rasmussen, A. P., & de Vries, C. P. 2006, ApJ, 652, 189

Katz, N., Weinberg, D. H., Hernquist, L., & Miralda-Escude, J. 1996, ApJ, 457, L57+

Kirkman, D., Tytler, D., Suzuki, N., O'Meara, J. M., & Lubin, D. 2003, ApJS, 149, 1

Kohler, K., & Gnedin, N. Y. 2007, ApJ, 655, 685

Komatsu, E., Smith, K. M., Dunkley, J., Bennett, C. L., Gold, B., Hinshaw, G., Jarosik, N., Larson, D., Nolta, M. R., Page, L., Spergel, D. N., Halpern, M., Hill, R. S., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2011, ApJS, 192, 18

Krauss, L. M., & Turner, M. S. 1995, Gen. Relativ. Gravit., 27, 1137

Lemaître, G. 1931, MNRAS, 91, 483

Mandelbaum, et al. (2003), Monthly Notices of the Royal Astronomical Society, Volume 344, Issue 3, pp. 776–788

Mathur, S., Weinberg, D. H., & Chen, X. 2003, ApJ, 582, 82

Miralda-Escude, J., Cen, R., Ostriker, J. P., & Rauch, M. 1996, ApJ, 471, 582

Mushotzky, R. F., & Loewenstein, M. 1997, ApJ, 481, L63+

Nicastro, F., Mathur, S., Elvis, M., Drake, J., Fang, T., Fruscione, A., Krongold, Y., Marshall, H., Williams, R., & Zezas, A. 2005, Nature, 433, 495

Ohashi, T., Ishisaki, Y., Ezoe, Y., Sasaki, S., Kawahara, H., Mitsuda, K., Yamasaki, N. Y., Takei, Y., Ishida, M., Tawara, Y., Sakurai, I., Furuzawa, A., Suto, Y., Yoshikawa, K., Kawai, N., Fujimoto, R., Tsuru, T. G., Matsushita, K., & Kitayama, T. 2010, in SPIE Conf. Ser., Vol. 7732 (Bellingham, WA: SPIE)

Peacock, J. 1992, Nature, 355, 203

Peebles, P. J. E. 1993, in, Principles of Physical Cosmology, ed. P. J. E. Peebles (Princeton: Princeton University Press)

Penton, S. V., Stocke, J. T., & Shull, J. M. 2004, ApJS, 152, 29

Péroux, C., McMahon, R. G., Storrie-Lombardi, L. J., & Irwin, M. J. 2003, MNRAS, 346, 1103

Pettini, M. 1999, in Chemical Evolution from Zero to High Redshift, ed. J. R. Walsh & M. R. Rosa (Berlin/New York: Springer), 233+

Pettini, M., Madau, P., Bolte, M., Prochaska, J. X., Ellison, S. L., & Fan, X. 2003, ApJ, 594, 695

Pettini, M., Smith, L. J., King, D. L., & Hunstead, R. W. 1997, ApJ, 486, 665

Pettini, M., Zych, B. J., Murphy, M. T., Lewis, A., & Steidel, C. C. 2008, MNRAS, 391, 1499

Pontzen, A., Governato, F., Pettini, M., Booth, C. M., Stinson, G., Wadsley, J., Brooks, A., Quinn, T., & Haehnelt, M. 2008, MNRAS, 390, 1349

Prochaska, J. X., Gawiser, E., Wolfe, A. M., Castro, S., & Djorgovski, S. G. 2003, ApJ, 595, L9

Prochaska, J. X., Herbert-Fort, S., & Wolfe, A. M. 2005, ApJ, 635, 123

Prochaska, J. X., O'Meara, J. M., & Worseck, G. 2010, ApJ, 718, 392

Prochaska, J. X., & Wolfe, A. M. 1997, ApJ, 487, 73

Prochaska, J. X., & Wolfe, A. M. 2009, ApJ, 696, 1543

Rasmussen, A. P., Kahn, S. M., Paerels, F., Herder, J. W. d., Kaastra, J., & de Vries, C. 2007, ApJ, 656, 129

Rauch, M. 1998, ARA&A, 36, 267

Rauch, M., Haehnelt, M., Bunker, A., Becker, G., Marleau, F., Graham, J., Cristiani, S., Jarvis, M., Lacey, C., Morris, S., Peroux, C., Röttgering, H., & Theuns, T. 2008, ApJ, 681, 856

Rauch, M., Miralda-Escude, J., Sargent, W. L. W., Barlow, T. A., Weinberg, D. H., Hernquist, L., Katz, N., Cen, R., & Ostriker, J. P. 1997, ApJ, 489, 7

Ryan-Weber, E. V., Pettini, M., & Madau, P. 2006, MNRAS, 371, L78

Ryan-Weber, E. V., Pettini, M., Madau, P., & Zych, B. J. 2009, MNRAS, 395, 1476

Schaye, J., Aguirre, A., Kim, T., Theuns, T., Rauch, M., & Sargent, W. L. W. 2003, ApJ, 596, 768

Schramm, D. N., & Turner, M. S. 1998, Rev. Mod. Phys., 70, 303

Seljak, U., Makarov, A., McDonald, P., Anderson, S. F., Bahcall, N. A., Brinkmann, J., Burles, S., Cen, R., Doi, M., Gunn, J. E., Ivezić, Ž., Kent, S., Loveday, J., Lupton, R. H., Munn, J. A., Nichol, R. C., Ostriker, J. P., Schlegel, D. J., Schneider, D. P., Tegmark, M., Berk, D. E., Weinberg, D. H., & York, D. G. 2005, Phys. Rev. D, 71, 103515

Shull, 2009, Future Directions in Ultraviolet Spectroscopy: A Conference Inspired by the Accomplishments of the Far Ultraviolet Spectroscopic Explorer Mission. AIP Conference Proceedings, Volume 1135, pp. 301–308

Simcoe, R. A. 2006, ApJ, 653, 977

Simcoe, R. A., Sargent, W. L. W., & Rauch, M. 2002, ApJ, 578, 737

Simcoe, R. A., Sargent, W. L. W., & Rauch, M. 2004, ApJ, 606, 92

Songaila, A. 2001, ApJ, 561, L153

Songaila, A. 2005, AJ, 130, 1996

Spergel, D. N., Verde, L., Peiris, H. V., Komatsu, E., Nolta, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2003, ApJS, 148, 175

Steenbrugge, K. C., Nicastro, F., & Elvis, M. 2006, in The X-ray Universe 2005, ESA Special Publication, Vol. 604, ed. A. Wilson (Noordwijk: ESA Publications Division, ESTEC), 751+

Storrie-Lombardi, L. J., & Wolfe, A. M. 2000, ApJ, 543, 552

Storrie-Lombardi, L. J., McMahon, R. G., Irwin, M. J., & Hazard, C. 1994, ApJ, 427, L13

Takei, Y., Henry, J. P., Finoguenov, A., Mitsuda, K., Tamura, T., Fujimoto, R., & Briel, U. G. 2007, ApJ, 655, 831

Tescari, E., Viel, M., Tornatore, L., & Borgani, S. 2009, MNRAS, 397, 411

Tripp, T. M., Savage, B. D., & Jenkins, E. B. 2000, ApJ, 534, L1

Weinberg, D. H., Miralda-Escude, J., Hernquist, L., & Katz, N. 1997, ApJ, 490, 564

Womble, et al. 1996, Cold Gas at High Redshift. Proceedings of a workshop celebrating the 25th anniversary of the Westerbork Synthesis Radio Telescope, held in Hoogeven, 28–30 August, 1995, The Netherlands, (Dordrecht: Kluwer Academic Publishers), c1996, ed. M.N. Bremer, & N. Malcolm (Astrophysics and Space Science Library), Vol. 206, p.249

Zhang, Y., Anninos, P., & Norman, M. L. 1995, Bull. Am. Astron. Soc., 27, 1412

Zwaan, M. A., & Prochaska, J. X. 2006, ApJ, 643, 675

13 Cosmic Microwave Background

John Mather[1] · *Gary Hinshaw*[2] · *Lyman Page*[3]
[1]Astrophysics Science Division, NASA/GSFC Code 443, Observational Cosmology, Greenbelt, MD, USA
[2]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada
[3]Department of Physics, Princeton University, Princeton, NJ, USA

1 *Introduction* 611
1.1 Outline 611
1.2 Modern View 612
1.3 Prediction and Discovery 613
1.4 Blackbody Form and Dominance 614
1.5 Celestial Emission at CMB Frequencies 616
1.6 Energy Release, Anisotropy, Standard Model, and Polarization 616

2 *CMB Spectrum Theory and Measurements* 620
2.1 Major Questions and Spectrum Distortions 620
2.2 Bose-Einstein μ Distortion 620
2.3 Compton y Distortion 621
2.4 Recombination Details 621
2.5 Free-Free Distortion 622
2.6 Particle Decay and Annihilating Particles 622
2.7 Alternatives to the Hot Big Bang 623
2.8 Tests of Cosmic Inflation: Silk Damping 624
2.9 Dark Energy 625
2.10 Other Processes 625
2.11 Reionization History and the X-ray Background 626
2.12 Far-IR Background Sources 626

3 *High-Precision Spectrum Measurements* 626
3.1 Minimizing Foregrounds 627
3.1.1 Earth and Atmosphere 628
3.1.2 Solar System 628
3.1.3 Galactic Electrons 629
3.1.4 Galactic Dust 629
3.1.5 Galactic Atoms, Ions, and Molecules 630
3.2 COBE FIRAS 630
3.2.1 FIRAS Design 630
3.2.2 FIRAS Detectors and Data Processing 632

T.D. Oswalt, W.C Keel (eds.), *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-94-007-5609-0_13, © Springer Science+Business Media Dordrecht 2013

3.2.3 FIRAS Calibration 632
3.2.4 Foreground Removal for FIRAS 634
3.2.5 Limits on CMB Spectrum Distortion 635
3.3 COBRA Rocket Experiment 635
3.4 ARCADE 2 636
3.4.1 ARCADE 2 Calibration 638
3.4.2 ARCADE 2 Results 638
3.5 TRIS 639
3.6 PIXIE 640
3.7 DARE 643
3.8 Ultimate Limits 643

4 Historical Overview of Temperature and Polarization Anisotropy 644
4.1 Anisotropy Searches Prior to COBE 644
4.2 From COBE to WMAP 646

5 WMAP 648
5.1 WMAP Sky Maps 649
5.2 WMAP Angular Power Spectrum 651
5.3 Beyond WMAP 654

6 The Standard Cosmological Model 655
6.1 The Origin of Structure 657
6.2 Geometry of the Universe 658
6.3 The Matter Content of the Universe 659
6.4 The Age of the Universe 660
6.5 Initial Conditions: The Inflationary Parameters 661
6.6 Gravitational Waves and CMB Polarization 663
6.7 Building on the Standard Model 664

7 Anisotropy and Polarization Measurement Frontiers 665
7.1 Large Angular Scale B-mode Experiments 666
7.2 Small-Scale Anisotropy, $\ell > 2{,}000$ 667
7.2.1 Small Angular Scale Polarization 668
7.2.2 Lensing of the CMB 670
7.2.3 Neutrinos 675
7.3 The Future 679

Acknowledgments 680

References 681

Abstract: The cosmic microwave background (CMB) radiation, the relic of the early phases of the expanding universe, is bright, full of information, and difficult to measure. Along with the recession of galaxies and the primordial nucleosynthesis, it is one of the strongest signs that the Hot Big Bang Model of the universe is correct. It is brightest around 2 mm wavelength, has a temperature of $T_{\mathrm{cmb}} = 2.72548 \pm 0.00057$ K, and has a blackbody spectrum within 50 parts per million. Its spatial fluctuations (around 0.01% on 1° scales) are possibly the relics of quantum mechanical processes in the early universe, modified by processes up to the decoupling at a redshift of about 1,000 (when the primordial plasma became mostly transparent). In the cold dark matter (DM) model with cosmic acceleration (ΛCDM), the fluctuation statistics are consistent with the model of inflation and can be used to determine other parameters within a few percent, including the Hubble constant, the Λ constant, the densities of baryonic and dark matter, and the primordial fluctuation amplitude and power spectrum slope. In addition, the polarization of the fluctuations reveals the epoch of reionization at a redshift approximately twice that determined from the Gunn-Peterson trough due to optically thick Lyman α absorption in QSO spectra. It is of historic importance, and a testament to the unity of theory and experiment, that we now have a standard model of cosmology that is consistent with all of the observations.

Current observational challenges include (1) improvement of the spectrum distortion measurements, especially at long wavelengths, where the measured background is unexpectedly bright; (2) the search for the B-mode polarization (the divergence-free part of the polarization map), arising from propagating gravitational waves; and (3) the extension of fluctuation measurements to smaller angular scales. Much more precise spectrum observations near 2 mm are likely and would test some very interesting theories. Current theoretical challenges include explanation of the dark matter and dark energy; understanding, estimating, and removing the interference of foreground sources that limit the measurements of the CMB; detailed understanding of the influence of nonequilibrium processes on the decoupling and reionization phases; and searches for signs of the second order or exotic processes (e.g., isocurvature fluctuations, cosmic strings, non-Gaussian fluctuations). At this writing, we await the cosmological results of the Planck mission.

Keywords: Alpher, Anisotropy, ARCADE, Big Bang Theory, Blackbody, Bose-Einstein, CMB, COBE, Cold dark matter, Compton distortion, Cosmic microwave background radiation, Decoupling, Dicke, DMR, FIRAS, Foregrounds, Galactic emission, Herman, Lensing, Penzias, PIXIE, Planck, Polarization, Silk damping, Spectrum distortion, Standard model, Steady State Theory, Sunyaev–Zel'dovich, Wilson, WMAP

1 Introduction

1.1 Outline

In this article we outline the importance and the modern view of the CMB, its prediction and discovery, and the reason for its blackbody form. After a summary introduction we discuss, in section 2, the theory and measurement of the CMB spectrum and its distortions. We then discuss the theory and measurements of the CMB spectrum and its distortions, the standard μ and y distortions, the details of recombination, the effects of particle decay and annihilation, alternatives to the Hot Big Bang, tests of cosmic inflation through the Silk damping effects, the effects of the dark energy, other processes, the reionization history, and far IR sources.

In section 3 we discuss high-precision spectrum measurements, outlining the techniques of differential comparison with reference blackbodies, as illustrated by the COBE FIRAS instrument, the COBRA rocket-borne instrument, the ARCADE 2 balloon-borne instrument, and the TRIS ground-based long wavelength measurements. We describe the PIXIE proposed high-precision spectropolarimeter and briefly discuss the DARE mission to search for the effects of the redshifted 21 cm hydrogen line.

The treatment of anisotropy and polarization measurements is somewhat different because there is a companion article describing the instrumentation by Hanany et al. (2012). We briefly review the history of temperature and polarization measurements and discuss the WMAP results, the standard cosmological model, the origin of structure, the geometry of the universe, the matter content of the universe, the age of the universe, the initial conditions from inflation, and parameters beyond the standard model. We discuss the anisotropy and polarization measurement frontiers, the search for non-Gaussianity in the fluctuations, and the search for large angular scale B-mode polarization, small scale anisotropy and polarization, the effects of lensing of the CMB, and the effects of neutrinos on the CMB.

1.2 Modern View

The cosmic microwave background radiation is the measurable relic of nature's greatest particle accelerator, the hot Big Bang, in which the temperatures and densities were so high that all particle species, and possibly all their collective oscillation modes, were in local thermal equilibrium. These particles must include quarks, leptons including the cosmic neutrino background, the Higgs boson, dark matter, supersymmetric particles if they exist, and the carriers of the four known forces: weak and strong nuclear forces, electromagnetism, and gravitation. If the Standard Model of particle physics is correct, then the weak and strong nuclear forces and electromagnetism are all unified in a single description, and all have comparable strength at high-enough temperatures and densities. Pushing back even farther, we can imagine an inflationary period in which a false vacuum filled with a scalar or other fields would decay into the true vacuum we observe today, along with particles and exponential expansion. And perhaps there was an era of quantum gravity in which gravitation was unified with the other three forces, and space and time were themselves quantum phenomena. But as the universe cooled, symmetries were broken many times as structures developed from the primordial material, antimatter was annihilated, and energy liberated from those phase transitions has been added to the electromagnetic fields, now observed at microwave frequencies. As a result, the CMB is now the dominant electromagnetic radiation field in the universe, and its photons far outnumber the baryons. The only other free particles with comparable densities are the unobserved cosmic background neutrinos, and potentially the dark matter particles, whatever they may be.

Of course there are innumerable virtual particles and vacuum fluctuations, whose influence can be measured in the Casimir effect, but whose meaning is not fully appreciated. Also, there may be holographic quantum fluctuations of space-time itself, but that is another topic. And curiously enough, since photons in vacuum are massless, the proper time for their trajectories from the Big Bang to our receivers is exactly zero. (On the other hand, the idea of the photon trajectory is itself a bit fuzzy, since electromagnetic fields are not billiard balls, and there are plasma interactions.)

In the homogeneous and isotropic expanding universe model, the CMB is approximately isotropic in a "preferred rest frame" at each point in space time. More precisely, the dipole term, that is, the lowest spherical harmonic of the distribution of the CMB temperature, is zero. That means that the velocity of the Earth relative to a large sample of the early universe can be observed and explained, but no more fundamental consequence has been recognized. It is also interesting that a substantial velocity of the observer relative to the CMB would change the angular scale of the features, owing to the aberration of light.

1.3 Prediction and Discovery

Prediction. The expanding universe was predicted by Friedman (1922) based on the cosmological equations of Einstein (1917), and independently by Lemaître (1927) who also estimated the Hubble constant from the known observations, and measured by Hubble (1929) using Cepheid variables as distance indicators. But it might or might not have been hot in the beginning, and the Steady State Theory of Hoyle (1948) and Bondi and Gold (1948) requiring replenishment by matter creation might have been correct, so the prediction and discovery of the CMB were hugely important for cosmology. When the expanding universe was first recognized, the distance measurements were seriously incorrect. Hubble's Cepheid variables in the Milky Way were a different type from the ones he found in other galaxies, leading to an expansion age that was significantly less than the ages of stars and even the Solar System, and casting doubt on the whole concept of the Big Bang. The discovery of the CMB did not quite erase all doubt about the hot Big Bang, as there was still the possibility that either a cold Big Bang or a steady state universe would produce starlight that could be absorbed and reemitted by dust grains to fill the universe with microwave background radiation. Also, even into the 1990s, there were questions about the expansion age of the universe relative to the oldest stars. This issue was not resolved until the launch and repair of the Hubble Space Telescope, which enabled more precise measurements of the Hubble constant, detection of the acceleration of the expansion, and better understanding of stellar ages.

Early Unrecognized CMB Measurements. The first known observation that can now be interpreted as a measurement of the CMB was reported very briefly by McKellar (1941), and mentioned by Herzberg (1950), a textbook with a statement that it had only limited significance. The observation of the ultraviolet absorption lines of interstellar CN molecules yielded their rotational excitation temperature, and of course one could imagine many possible ways that the molecules could be excited. Hence, the observation was not pursued at the time.

Alpher and Herman Prediction of Temperature. Alpher and Herman (1948), working with G. Gamow, estimated the temperature of the CMB at 5 K; later they estimated 28 K. They (Alpher, personal communication) tried to convince observers to go look for it, but at that time no serious effort was made, and in any case it would have been extremely difficult with the technology available then. J. Weber wanted to try but was told directly that the measurement was impossible (Weber, personal communication). The Alpher and Herman papers did not emphasize the predicted spectrum of the CMB or compare the spectrum with foreground sources or discuss how it might be detected. Later, radio astronomers and engineers made a number of measurements of the temperature of the dark sky and gave evidence that it was not in their instruments, but none of them were recognized as strong enough evidence or sufficiently surprising to the observers to command attention.

Discovery. The CMB was finally discovered at Bell Telephone Labs by Penzias and Wilson (1965), who were not looking for it but were astronomers testing a new, sensitive, and stable microwave receiver and horn antenna, working at 7.35 cm wavelength. As they were rechecking their measurement, they learned through B. Burke of a group at Princeton who were building equipment to measure the radiation. The Princeton group had been thinking about the bouncing universe, which would be filled with photons left over from previous expansion/contraction cycles. The Bell Labs discovery was published simultaneously with the Princeton interpretation by Dicke et al. (1965) and was front-page news in the NY Times (May 21, 1965). The Princeton group (Roll and Wilkinson 1966) soon completed their measurement at 3.2 cm wavelength, confirming or at least making it plausible that the CMB has the blackbody spectrum required by the hot Big Bang idea.

For a remarkable historical summary of the discovery of the CMB and its properties, the book "Finding the Big Bang" by Peebles et al. (2009) gives the human side of this field as well as an excellent tutorial on the technical aspects.

1.4 Blackbody Form and Dominance

Prediction of Blackbody Form. The idea that the CMB is the remnant of a hot equilibrium phase implies that the spectrum must be very close to a blackbody spectrum, although when examined very closely there must be tiny differences due to the cooling of matter below the CMB temperature (Chluba and Sunyaev 2012a). The blackbody form is preserved exactly through the history of the expanding universe, according to the following simple argument: Imagine a box containing primordial CMB, and imagine that the box expands with the homogeneous and isotropic expanding universe. Then photons crossing the walls of the box are in detailed balance, so we may now replace the imaginary box with a real box of moving mirrors that move with the expanding universe. Within this mirror box, we may represent the electromagnetic field as quanta occupying the spatial and polarization modes of the box. As the box expands, these modes expand adiabatically, and the occupation numbers do not change. The energy of each quantum diminishes as the box expands as well. In combination, these factors imply that the Planck function description of the blackbody is preserved, and the occupation number of each mode is just $1/(e^x - 1)$, where $x = h\nu/kT$. Also, the temperature of the CMB is inversely proportional to the expansion factor of the universe; conversely, $T_{\rm cmb} = T_0(1+z)$, where $T_{\rm cmb}$ is the temperature of the CMB at a time in the past, T_0 is its temperature now, and z is the redshift corresponding to the time in the past. This dependence has been confirmed by observations of the excitation temperature of cyanogen (CN) and other molecules and ions seen in absorption against quasars, and through observations of the Sunyaev-Zel'dovich effect in distant clusters of galaxies.

Dominance and Perfect Spectrum. On a cosmic scale the CMB is extraordinarily bright, even though it has been difficult to measure. Its brightness (σT^4, with $T = 2.72548\,\rm K$) is $3.129\,\mu\rm W/m^2$. The prediction of the blackbody form is very robust because the photons outnumber the baryons by nine orders of magnitude, and it is very difficult to conceive of any way in which they could have modified the CMB spectrum very much. This difficulty is a matter of perspective and scale – there are many kinds of proposed exotic processes such as explosive events that could have modified the spectrum, as well as four processes that are expected to occur: (i) acoustic damping, (ii) cooling of photons by adiabatically cooling matter, (iii) recombination radiation, and (iv) depending on the mass of the Dark Matter (DM) particle, also DM annihilation. All of these processes are explained in Chluba and Sunyaev (2012a). The current success of the ΛCDM model has shifted focus from the more radical of these ideas.

Spectrum Turnover. Early measurements of the CMB were made at long wavelengths (more than a few mm) where the Planck function is close to a power law, that might also occur from a nonthermal process, so it was important to observe at shorter wavelengths. Many additional measurements of the CMB spectrum were made from the ground, balloons, rockets, and interstellar molecules, eventually confirming the blackbody turnover at short wavelengths. Until the flight of the Cosmic Background Explorer (COBE) satellite in 1989, reported by Mather et al. (1990), there was evidence that the spectrum was not exactly blackbody, as reported, for example, by Matsumoto et al. (1988), suggesting excess brightness at short wavelengths, but only with limited accuracy. The COBE results were quickly confirmed by the COBRA experiment of Gush et al. (1990). Reasons for the difficulties include: the CMB is faint relative to our 300 K local environment, it is nearly isotropic so that measurements must be absolutely calibrated, receiver sensitivity was barely adequate, atmospheric emission is strong, galactic emission from electrons (free-free scattering and synchrotron) is bright at long wavelengths, galactic dust is bright at short wavelengths, and instruments operated in air cannot be cooled to temperatures comparable to the CMB so that instrument self-emission is strong and absolute calibration is difficult. ➲ *Figure 13-1* shows the original COBE-FIRAS spectrum that showed that the COBE was working well. It is now the iconic figure even though the error bars have been reduced to 50 parts per million by Fixsen et al. (1996); see below for details.

Where Does the Energy Go? The energy density of the CMB and other constituents of the universe decline with the expansion, so where does that energy go? There are a few surprises. First, energy alone is not a conserved quantity in relativity. In special relativity, it is one component of a four-vector, and mass and energy can be interconverted according to $E = mc^2$.

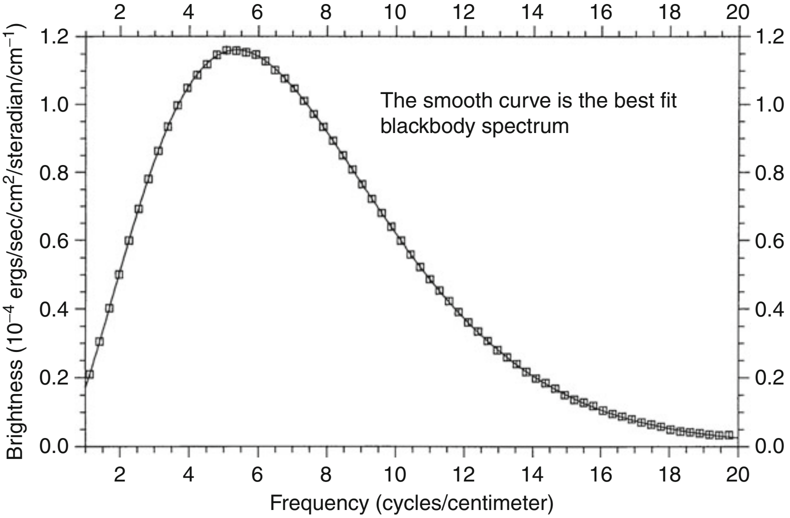


Fig. 13-1

Preliminary spectrum of the cosmic microwave background from the FIRAS instrument at the north Galactic pole, compared to a blackbody. *Boxes* are measured points and show the assumed 1% error band. The units for the vertical axis are 10^{-4} ergs s^{-1} cm^{-2} sr^{-1} cm (From Mather et al. (1990))

In general relativity it is only one component of a stress-energy tensor, and in both cases its numerical value changes according to the velocity of the coordinate system in which it is measured. But there are still local conservation laws. Second, the "universe" is not a closed, finite system. In the expanding mirror box described above, the photons inside the box do work on the walls of the box, but when we remove the mirrors and replace the photons with others coming from the other side, there is nothing receiving the work that is being done on the imaginary walls. We conclude that we should trust the differential equations of general relativity but not the simplifications from analogies. There was a reason that cosmology could not be completed with Newtonian mechanics and nineteenth-century thermodynamics.

1.5 Celestial Emission at CMB Frequencies

➋ *Figure 13-2* shows the antenna temperature of the sky from 1 to 1,000 GHz for a region at a galactic latitude of roughly 20°, though the levels one measures can be different by an order of magnitude depending on galactic longitude. The antenna temperature of a gray body is $T_{\text{ant}} = \epsilon T x/(e^x - 1)$, where T is the physical temperature, ϵ is the emissivity, and $x = h\nu/kT$. Ignoring emission from the atmosphere, synchrotron emission dominates celestial emission at the low-frequency end, and dust emission dominates at high frequencies. The basic picture in ➋ *Fig. 13-2* has remained the same for over 30 years (Weiss 1980), though over the past decade, there has been increasing evidence for a new component of celestial emission in the 30 GHz region (e.g., Kogut et al. 1996; de Oliveira-Costa et al. 1997; Leitch et al. 1997). This new component is spatially correlated with dust emission. It has been identified with emission by tiny grains of dust that are spun up to GHz rotation rates by a variety of mechanisms, so-called "spinning dust," though other emission mechanisms may contribute to or produce the signal (Draine and Lazarian 1998, 1999). Understanding this emission source is an active area of investigation.

1.6 Energy Release, Anisotropy, Standard Model, and Polarization

Limits on Early Energy Release. If the CMB spectrum does not match a blackbody form, then significant energy release must have occurred to change it. When the universe was about 1 year old (redshift about 2×10^6), the double-photon Compton scattering processes that create and destroy photons effectively ceased, but multiple Compton scatterings that equilibrate energy between wavelengths were still operating. Hence, if energy were added or removed from the CMB, for instance by the decay of some dark matter particle, then the CMB could have a spectrum with a dimensionless chemical potential μ, and the photon occupation number would equilibrate to the form $1/(e^{x+\mu} - 1)$. (This form is only valid at high frequencies; at low frequencies, μ has to be a function of frequency.) When the universe cooled sufficiently to stop even this equilibration process, it became possible that we would observe a mix of blackbodies at different temperatures, either from a simple mixing or from energy added by Compton scattering from hot electrons. This is described by the Kompaneets parameter y, and the first serious limits were set by the COBE FIRAS instrument. Less than 0.01% of the CMB energy was added after the first year.

Events at and After Decoupling. About 400,000 years later, the universe became fairly quickly (over a period of about 100,000 years) transparent when temperatures reached around

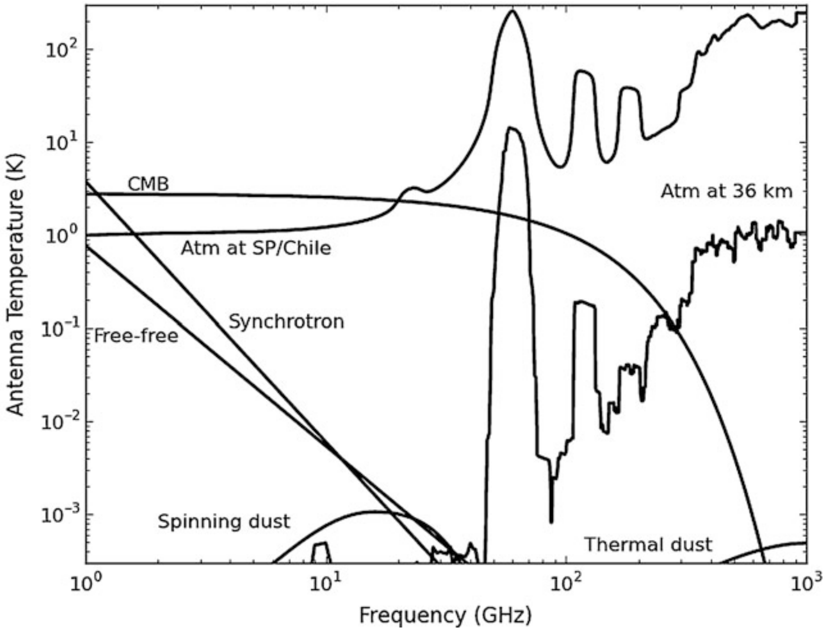


Fig. 13-2

The antenna temperature from 1 to 1,000 GHz for a region of sky near a galactic latitude of roughly 20°. The flat part of the CMB spectrum, roughly below 30 GHz, is called the Rayleigh-Jeans portion. A Rayleigh-Jeans source with frequency-independent emissivity is a *horizontal line* on this plot. The synchrotron emission is from cosmic ray electrons orbiting in galactic magnetic fields and is polarized. Free-free emission is from galactic electrons' "braking radiation" (bremsstrahlung) and is not polarized. The amplitude of the spinning dust is not well known. This particular model comes from Ali-Haïmoud et al. (2009). The standard spinning dust emission is not appreciably polarized. The atmospheric models are based on the ATM code (Pardo et al. 2001) and are for a zenith angle of 45°. The South Pole/Atacama (Chile) spectrum is based on a precipitable water vapor of 0.5 mm. The difference between the two sites is inconsequential for this plot. The atmospheric spectra have been averaged over a 10% bandwidth. The pair of lines at 60 and 120 GHz are the oxygen doublet. The lines at 19 and 180 GHz are vibrational water lines. The finer scale features are from ozone

3,000 K and electrons were bound to atomic nuclei. This moment is known as the decoupling. We observe the map of the CMB predominantly as it was when it was last scattered in our direction. Little effect on the spectrum can be produced by the details of the reactions because as noted above, the photons outnumber the baryons by an enormous factor. But after decoupling, Compton drag on the residual ionization of the baryonic material still limits its ability to move, and, in addition, the baryons can cool adiabatically to have a temperature less than that of the CMB. These effects have tremendous importance to the formation of stars and galaxies, even if we cannot see much effect on the CMB. (Note that although it is often said that we observe the universe at the decoupling, the CMB spectrum is determined by and responds to events back to year one.)

Isotropy. The first test for the cosmic nature of the CMB was that it is isotropic (the same brightness in all directions). If its brightness showed any correlation with known objects such as the weather or the ecliptic plane or the galactic plane or nearby galaxies or clusters, then it would not be cosmic. As it happens, the foregrounds are bright enough to matter, but can be measured at other wavelengths and modeled and extrapolated to find a residual background. For many years, the only results on anisotropy were upper limits, but eventually the Doppler shift due to the Earth's motion relative to the cosmos was measured as a significant dipole. The Doppler-shifted blackbody is still a blackbody but at a modified temperature. We now know the velocity of the Solar System relative to the cosmos as $v = 369.0 \pm 0.9\,\mathrm{km\,s^{-1}}$ from COBE and WMAP (Hinshaw et al. 2009). The velocity produces a CMB temperature distribution that is to first order $T = T_0(1 + (v/c)\cos\theta)$ where, v is the velocity of motion, and θ is the angle between the observed direction and the direction of motion. The measured velocity is the vector sum of the instrument's velocity around the Earth and the Sun, the Sun's velocity around the center of the Milky Way galaxy, and the Milky Way's velocity relative to the rest of the universe, or more precisely the slice of space-time when the universe became transparent to the photons now reaching our detectors. The masses of nearby galaxies, acting through gravitational attraction over cosmic time, are thought to be enough to explain the motion of the Milky Way and hence the vector sum.

Higher Order Anisotropy. As it happens, the decoupling was soon after a time when baryonic matter was beginning to move under the influence of gravitational forces, now that it was no longer so strongly tied by Compton scattering to the CMB radiation field, and the attenuation of the radiation temperature diminished the gravitational importance and the pressure of the radiation field itself. (Dark matter was free to move much sooner.) On angular scales greater than 7°, the major feature is the Sachs–Wolfe effect (Sachs and Wolfe 1967), in which some regions of the universe are more dense than others, and photons leaving the dense regions suffer more gravitational redshift than others. It was predicted by Harrison (1970), Peebles and Yu (1970), and Zel'dovich (1972) on very general grounds that the primordial fluctuations should have a scale-free power spectrum, with equal fluctuation amplitudes on all spatial scales. This prediction has been confirmed to excellent precision and extended by the WMAP team and by other anisotropy measurements at small angular scales. In addition, some forms of inflation theory say that the spectral index should not be exactly unity as predicted by Harrison, Peebles and Yu, and Zel'dovich, but a few percent smaller; this has also been has been measured as discussed below.

For smaller angular scales, a detailed analysis of coupled fluids acting before and after the decoupling is necessary. The fluids include the CMB, the cosmic neutrino background, the baryonic matter, the dark matter (cold and/or warm), and the dark energy (affecting the recent expansion history). In a remarkable accomplishment, cosmologists agree very well on the equations to be solved and the methods to be used; this is possible because all the motions are small and the complexity of the modern universe has not yet developed. There is now a "standard model" of cosmology based on cold dark matter with an Einstein Λ constant that matches all of the observations of anisotropy on all measured scales from the quadrupole term (90° angular scale) up to multipole orders of thousands (arcminute scales). This is true despite the tiny amplitude of the fluctuations: a part in 10^5 on 7° scales, a part in 10^4 on 1° scales.

The remarkable feature found by observations and matched by theory is that there is a preferred angular scale for the fluctuations, called the "acoustic peak," at about 1° scale (spherical harmonic order about 200). This is effectively the observed size of the event horizon (and the age of the universe then) at the time of decoupling, and as matter feels the gravitational fields of primordial perturbations, it begins to move at the decoupling time. More precisely, the event horizon (at the speed of light) is about 1.2°, and the acoustic horizon is about 0.6°, smaller because the speed of sound then is about half the speed of light. This acoustic horizon size at decoupling provides a physical scale that is imprinted on the matter distribution and preserved as the universe expands. It can be detected in the galaxy-galaxy correlation function and is the basis for the baryon acoustic oscillation method of measuring the cosmic acceleration, since the apparent size of a physical marker can be measured as a function of redshift. At smaller angular scales, there are approximate harmonics of the fundamental oscillation frequency, and the small-scale oscillations lose amplitude because of photon diffusion, an effect called Silk damping.

Propagation. Note that the intergalactic medium is not perfectly transparent; after decoupling, there is still residual ionization, and then after the first bright UV sources arise, the universe becomes reionized. The optical depth of the intergalactic medium after decoupling has been measured by the correlation of anisotropies and polarization on relatively large angular scales using the WMAP data. According to the WMAP7 data set (Jarosik et al. 2011), the optical depth from here to the decoupling is about $\tau = 0.088 \pm 0.015$, and the reionization redshift was $z = 10.5 \pm 1.2$, assuming the reionization happened quickly.

Lensing. A somewhat surprising result of general relativity is that the distant universe is not where it seems to be, but may be arcminutes away, due to the effect of gravitational lensing of intervening clusters and superclusters of galaxies. The Millennium Simulation yielded a mean deflection angle of $2.4'$ (Carbone et al. 2009). The lensing preserves surface brightness, so might naïvely be expected to have no effect on the CMB and its anisotropy, but this is untrue because the lensing changes the observed angular scales of the fluctuations, magnifying some and shrinking others. Hence, the lensing can be detected statistically against the random fluctuation field of the CMB and used to measure parameters of the mass distribution of the universe. Lensing would not directly affect the polarization of a CMB photon much (because the deflection angles are small), but it does affect the spatial map of the polarization field and sets a limit on the measurement of primordial gravitational waves (see below).

Standard Model. The results of these observations, interpreted through the standard model, have produced a total transformation of the subject of cosmology, from highly speculative to highly quantitative. The parameters of the standard model can be measured with accuracies of the order of a few percent or better and are in agreement with measurements of the accelerating universe obtained in other ways (supernova distance scale, baryon acoustic oscillations, and clustering). On the other hand, some writers believe that warm dark matter may also be required, rather than or in addition to cold. Although the CMB observations are matched essentially perfectly by ΛCDM, the populations and spatial distributions of dwarf galaxies may not match the models so well. This is not simple to model or to observe, as all the complexities of star formation, stellar winds, galactic growth and evolution, black holes, and AGN can influence the comparison of observations with numerical simulations.

Polarization. The new and exciting challenge for CMB observers is to measure the polarization of the CMB. Some large-scale polarizations and correlations with the intensity anisotropy have already been observed by the WMAP team and interpreted to measure the ionization history after the decoupling and the onset of reionization, extending the direct measurements of quasar absorption lines. But the current challenge is to measure the effects of primordial gravitational waves. If such waves were in equipartition equilibrium with other fluctuation events in the early universe, then there should be a statistical signature left. In analogy with electromagnetic fields, the observed polarization vectors of the CMB can be broken down into a curl part with no divergence (B mode) and a divergence part with no curl (E mode). Primordial gravitational waves, unlike other cosmological perturbations, produce equal amounts of E and B modes. The mode of action is that the gravitational waves, still propagating at the epoch of decoupling, would be stretching and squeezing the primordial fluid so that there is a quadrupole intensity anisotropy at the time of the decoupling. This quadrupole anisotropy, incident on the electrons at last scattering, will result in polarization vector, proportional to two of the five components of the quadrupole term then. We then observe the polarization as a map of the spatial variation of that quadrupole. The predicted signal is very small. Nevertheless, calculation shows that the signal can be measurable, and hundreds of people are working on roughly ten different projects to do it. We cover this topic in more detail below.

2 CMB Spectrum Theory and Measurements

2.1 Major Questions and Spectrum Distortions

In this section we describe three predicted forms of distortion of the CMB spectrum and several effects that could alter the spectrum from the initially perfect pressure-cooker blackbody form. Did the universe really start with a hot Big Bang? Is there any effect of cosmic inflation on the spectrum of the CMB? Did some kind of matter decay or other energy source add energy to the CMB, in such a way that the effect could be seen today? Were there exotic processes like cosmic strings, explosive events, or abundant black holes in the early universe that could modify the CMB spectrum? When all the details about reactions and radiation transfer around the recombination are included, what difference do they make to the general history? Are there observable consequences of the atomic recombination sequences (H, He, He^+, Li) or molecules (LiH, H_2)? How does the CMB interact with the hyperfine structure of hydrogen, and what could redshifted 21 cm hydrogen emission and absorption tell us? What might make the unresolved excess background radiation seen at cm wavelengths? What is the history of structure formation? The history of reionization? What are the foreground sources, and how can we see past them?

2.2 Bose-Einstein μ Distortion

Wright et al. (1994) and Kogut et al. (2011) summarize the leading models for observable spectral distortions in the context of the FIRAS observations and the proposed PIXIE mission. Energy release within the first year of the expansion would simply be thermalized, without a change from the blackbody form, due to rapid Compton scattering and the double Compton process in which $e + \gamma \longleftrightarrow e + 2\gamma$.

But when the temperature drops sufficiently, this process becomes slow compared to the age of the universe. For redshifts from $z_y = 1.4 \times 10^5$ to $z_{\mathrm{th}} = 2 \times 10^6$, energy added to the CMB field will not be fully thermalized, because the photons are no longer freely created or destroyed, resulting in a Bose-Einstein distribution with a dimensionless chemical potential μ. Electrons still produce Doppler shifts when scattering the photons, and multiple scatterings are enough to produce a pseudoequilibrium form for the photon occupation number: $\eta = 1/(e^{x+\mu} - 1)$. Detailed calculations are reported by Burigana et al. (1991), Daly (1991), and Hu et al. (1994).

In this case the chemical potential $\mu = 1.4\Delta U/U$, where ΔU is the energy added to the CMB and U is its total energy (Sunyaev and Zel'dovich 1970b; Illarionov and Sunyaev 1975a, b).

For wide ranges of ν, it is convenient to let μ be a function of ν, $\mu(\nu)$. Note also that a negative μ could be produced if the electron cloud is colder than the CMB, which can happen as matter cools adiabatically below the CMB temperature. Negative μ would produce a singularity in the brightness temperature of the CMB at zero frequency, but this does not happen because the free-free opacity is large at low frequencies. Chluba and Sunyaev (2012b) show that negative μ would only be produced at redshift $z > 50{,}000$. Khatri et al. (2011) discussed this in depth and described it as a Bose-Einstein condensation.

2.3 Compton *y* Distortion

At lower redshifts than $z_{\mathrm{th}} = 1.4 \times 10^5$, multiple Compton scattering is too slow to produce an equilibrium spectrum. In that case, energy added to the CMB could retain some of its original spectrum. In particular, hot objects would produce additions to the short-wavelength end of the CMB spectrum, even though scattering would prevent observing them. In the absence of discrete hot objects or high energy photons, we consider the effects of scattering from warm electrons, heated above the CMB temperature by some energy source. Doppler shifts from Compton scattering by nonrelativistic electrons effectively produce a CMB spectrum that is a mixture of blackbodies at different temperatures. The resulting spectrum is parameterized by the Kompaneets y, where $y = \Delta U/4U$, and ΔU and U are the total energy release and the total energy, as before,

$$y = \int \frac{k(T_e - T_{\mathrm{cmb}})}{m_e c^2} d\tau, \tag{13.1}$$

where T_e is the electron temperature, T_{cmb} is the CMB temperature at the time, and τ is the optical depth to electron scattering. A modification is required if the electrons are relativistic (Wright 1979) because individual scatterings can shift photons far into the short wavelength Wien tail of the spectrum.

2.4 Recombination Details

An obvious question is whether the details of the recombination process at the time of decoupling can produce a measurable effect on the anisotropy or spectrum of the CMB. The short answer is no because there are 10^9 photons per baryon. However, the scattering optical depth of the hydrogen and helium transitions to and from the ground state is also extremely large, so there is a calculable delay in recombination, a slightly nonequilibrium distribution of populations of the levels, and a small trace of the hydrogen lines in the CMB spectrum. Peebles (1968a) discussed the Lyman α line emission, and as did Zel'dovich et al. (1969), Varshalovich and Khersonskii (1977), and Dubrovich and Stolyarov (1995). Dubrovich (1975) was the first to mention recombination lines at high n. An H atom in the $n = 2$ state faces a bottleneck in reaching the ground level: either it decays by a two-photon path or it emits a Lyman α photon that must escape an optically thick cloud. This trapping means that each atom can produce many residual photons resulting from transitions among the excited states, leading to an amplification of the spectrum distortions, but they are still too small to observe directly now.

On the other hand, at 2 GHz, the fractional distortion may be as large as 2×10^{-7}, and on the Wien end of the spectrum, the fractional distortion is large and affects the chemistry calculations at low redshfits (e.g., Switzer and Hirata 2005; Coppola et al. 2012a).

Detailed computer simulations have now been done, keeping track of the reaction rates and the populations of individual atomic, ionic, and molecular states of H, He, and Li. Key recent papers include Seager et al. (1999, 2011), Chluba and Sunyaev (2006), Rubiño-Martín et al. (2008), Sunyaev and Chluba (2009), and Chluba and Sunyaev (2012b). Switzer and Hirata (2008) discussed primordial helium recombination, Ali-Haïmoud and Hirata (2011) described the HyRec code for hydrogen and helium recombination, and Alizadeh and Hirata (2011) discuss the effects of possible molecular H_2 at the recombination era (z = 800 – 1,200). Coppola et al. (2012b) review prior work back to 1983 and discuss a detailed analysis of the reaction rates and effects of the formation of molecular hydrogen H_2. Peak production rates would occur around redshift $z = 100$, and the rotational-vibrational transitions would emit radiation peaking around 2 μm wavelength.

The predicted spectral distortion from all of these factors is very small but not necessarily unobservable, now that technology and concepts have improved. The Lyman series of spectral lines are the strongest but are broad, and the contrast against the continuum is of order 10^{-8}. The Lyman α line is redshifted to around 130 μm, where Galactic and zodiacal dust emission are bright, and atomic and molecular transitions in the interstellar medium of the Milky Way and external galaxies would confuse the search for the recombination line.

The electron gas cools adiabatically far below the CMB temperature when it is no longer strongly coupled (thermally and kinematically) to the CMB, but this event occurs around a redshift $z = 150$, and of course the optical depth is then also small. Galli et al. (2008) report that the "delayed recombination" has important effects in determining cosmological parameters from the WMAP data, but would be less important when smaller angular scale data are included. Peebles et al. (2000) demonstrated the effect for the first time and gave it the name.

2.5 Free-Free Distortion

At long wavelengths the free-free opacity after decoupling is significant and can either cool or heat the CMB depending on whether the electrons have been cooled or heated. Chluba and Sunyaev (2012b) included this effect in their numerical simulations. For the case of cooled electrons, they find spectral distortions similar to a negative y at high frequencies, created at late times, and a negative μ at low frequencies (1 GHz), created well before recombination ($z > 50{,}000$).

2.6 Particle Decay and Annihilating Particles

As noted by Wright et al. (1994), rare photons with $h\nu \gg m_e c^2/\tau_H$ would lose most of their energy by multiple Compton scattering and deliver their energy to the electron gas, which would then modify the spectrum of the CMB photons. In this formula, τ_H is the optical depth for electron scattering per Hubble time. Such photons must be rare; otherwise, we would have seen a large distortion of the CMB spectrum. The photons could come from the decay of rare particles or from the decay of common particles with small branching ratios. For instance, Fukugita and Kawasaki (1990) considered the decay of massive (20 keV) neutrinos at $z \approx 4{,}000$, which would produce a y distortion.

Some theories have posited that ordinary matter or dark matter may be slightly unstable. In the standard model of particle physics, baryon number is conserved, and the proton is the lightest baryon, so it must be stable. Observational limits on its lifetime are greater than about 10^{34} years. On the other hand, baryogenesis and the asymmetry between matter and antimatter are not understood, and some theories predict that protons are unstable. Also, since we do not know what particles comprise dark matter, perhaps it is the decay product of some other particles that might have had interesting lifetimes. A lifetime of one year corresponds to the time at which the CMB spectrum could begin to deviate from a blackbody form.

The dark matter might be neutralinos (supersymmetric partners of photons), which arise naturally in supersymmetry theories; there may be four neutralinos, of which only the lightest would be stable. The primordial supersymmetric particles, most of which are charged, could not be stable or we would still see them. Instead, annihilation of these charginos by their antiparticles would produce electromagnetic energy and neutralinos as dark matter. If dark matter particles are massive, then they would have cooled to low temperatures during the adiabatic expansion of the universe. If all the dark matter particles were produced in the first year of expansion, we would not see an effect on the CMB. But if they were produced later, and the energy could be coupled to the electromagnetic fields, then the CMB spectrum would be distorted.

Silk and Stebbins (1983), prior to the COBE launch, and later McDonald et al. (2001) reviewed some of the possibilities. According to de Vega and Sanchez (2010), CMB spectrum distortions would be most sensitive to neutralinos at a mass scale below 80 keV. Feng et al. (2003) computed the possible spectrum distortions as a function of particle lifetimes and electromagnetic energy release for two different particle models. They considered gravitinos (the supersymmetric partners of gravitons) as the dark matter and produced graphs illustrating the range of parameter space tested by CMB spectrum distortions. They show a range of the possible WIMP lifetime from 10^4 to 10^{12} s and a range of their energy release parameter ζ_{EM} from 10^{-12} to 10^{-7}.

Particles could also be annihilated by meeting their antiparticles, producing similar effects on the CMB spectrum. The CMB constraints on this process are already strong (Galli et al. 2008), as are those from the Fermi observatory, so predicted μ values are only a few times 10^{-9}.

2.7 Alternatives to the Hot Big Bang

Cold Big Bang. The current inflationary picture of the early universe provides a cold starting point, with a single scalar field rolling down into a potential well, and then radiating photons and other particles at extreme temperatures. But a more literal meaning, prior to the concept of inflation, held that the expanding universe was cold even after the particles were produced. If such a cold universe were unstable, then it could break up into clusters, galaxies, and stars, which would then liberate photons and dust, and then the dust would convert visible light into CMB. Lemaître (1931), immediately following the translation of his 1927 article, considered the instability of the early universe, as did Eddington and others. In those times there was also a theory that the cosmic rays are relics of the great explosion. Zel'dovich (1963) suggested that a zero-temperature early universe would undergo a phase transition and would then break up into condensations of planetary mass.

Layzer and Hively (1973) argued that the CMB could be produced in a cold Big Bang Model, if most of the matter in the universe were included in a population of stars at redshift 25–50.

The dust grains produced in these stars would thermalize the energy released by nuclear burning and supernova explosions. Aguirre (1999) reviewed the possibility of nucleosynthesis in such a cold Big Bang. He argues that the element abundances can be explained by a cold Big Bang and leaves open the question of whether the CMB could be astrophysically generated. His later paper Aguirre (2000) discusses the CMB and concludes that even with generous assumptions, the cold Big Bang idea is probably not correct.

Steady State Theory. The Steady State Theory held that the CMB (which had not been predicted by the theory) would be the accumulated, redshifted, thermalized radiation of all the stars. To achieve a steady state, it was necessary to posit an unobserved process of matter creation, to keep the mean density of the universe constant through time. The theory did not predict that the CMB should have a blackbody spectrum, and it also held that the temperature of the CMB should be independent of redshift, in conflict with observations of CN, CH, and [C II] lines in distant objects. Attempts to rescue the Steady State Theory in light of new data were increasingly ad hoc and required selective belief in certain observations and not others.

In both the Cold Big Bang and the Steady State theories, the optical depth and special optical properties of dust that would be needed to thermalize the CMB to the precise blackbody form would be quite unusual. Wright (1982) computed the properties of iron whiskers and outlined the tests to be made when a more precise spectrum of the CMB could be determined. But now that the CMB is known to have a precise blackbody form, the Cold Big Bang and the Steady State theories no longer match the data.

Of course, the radiation field produced by all the generations of stars, black holes, etc., and partially thermalized by dust does exist. It was measured at both near- and far-infrared wavelengths by the DIRBE team and is comparable in luminosity to all the visible classes of stars and galaxies. But it does not have a blackbody spectrum with emissivity near unity, as the CMB does.

2.8 Tests of Cosmic Inflation: Silk Damping

Do the primordial density fluctuations follow the power-law spectrum that has been tested so far by anisotropy measurements, and is its index different from the scale-invariant value? Below we discuss this in the context of direct measurements of the anisotropy. However, it is also possible to test this on small scales by searching for energy release at redshifts greater than 10^4, which would result in a distorted CMB spectrum. Primordial density fluctuations are frozen in place as the early universe inflates, and then come back into the horizon as expansion slows. As the smaller scale fluctuations come into view, the photons associated with them diffuse, the anisotropy is erased, and their energy can be converted to heat (Silk damping, Silk 1968; Sunyaev and Zel'dovich 1970b; Daly 1991; Hu et al. 1994). This effect is already taken into account in the prediction of anisotropies, but the effect on the spectrum would be too small to have been detected so far. The effect becomes a test of inflation because it is sensitive to primordial fluctuations on scales much smaller than those observable as CMB anisotropies. If those fluctuations were significantly stronger than the approximately scale-invariant prediction of inflation, then we would know from this measurement. Both y and μ distortions could be produced, depending on the physical scale at which the fluctuations deviate from the inflation prediction.

The limit set by the FIRAS instrument (below, Wright et al. (1994)) gives $|\mu| < 3.3 \times 10^{-4}$ (95% confidence). With this limit, the power law spectrum index of the fluctuations is less than

$1 + 6/7 \approx 1.9$. This result limits the fluctuations on scales much smaller than are accessible from the anisotropy measurements. PIXIE (see below) could do more than 4 orders of magnitude better, down to μ of 10^{-8}.

For later energy release, the first-order effect of Silk damping on the spectrum is again to mix together blackbodies at a range of temperatures, and $2y = \mathrm{var}(T)/T^2$, where var() is the variance. Note that because this is quadratic, the y resulting from the measured anisotropy of 10^{-4} is only 5×10^{-9} and not yet measurable. On the other hand, future precise measurements will have to account for this effect. For earlier energy release between z_y and z_{th}, the spectrum equilibrates to the Bose-Einstein form with $\mu = 2.8\,\mathrm{var}(T)/T^2$, also not yet detectable.

Chluba et al. (2012b) compute the effects of Silk damping of primordial acoustic waves, carrying the computations to second order in the perturbation amplitude and in the energy transfer. The acoustic energy can raise the temperature of the CMB (which we cannot distinguish from other effects) or it can cause both y and μ distortions. The μ distortion is particularly interesting because it occurs early. Future measurements of μ can set limits on the primordial fluctuations at scale sizes k from 50 to 10^4 Mpc^{-1}, even though these fluctuations are completely erased from the anisotropy and baryon distributions. Chluba et al. (2012a) show explicitly how powerful measurements of the CMB spectrum could be and that the COBE-FIRAS limit is already stronger than the primordial black hole limits.

2.9 Dark Energy

The cosmic acceleration has no predicted direct effect on the CMB except through modifying the expansion itself, and the CMB blackbody form is preserved. On the other hand, the details of the expansion history are critically important to the calculation of the anisotropy and its relation to present-day large-scale structure. The basic Sachs–Wolfe effect produces the primary anisotropies observed on large angular scales; dense regions cause gravitational redshifts of photons leaving them at decoupling. The integrated Sachs–Wolfe effect refers to the effect of the changing depth of the gravitational potential wells as the universe expands and dilutes the material and depends on whether the source of the gravitational potential includes radiation as well as matter.

2.10 Other Processes

Wright et al. (1994) summarize other exotic processes that might have occurred. Ostriker and Cowie (1981) considered an explosive scenario for galaxy formation; this is now ruled out as a general explanation. Gnedin and Ostriker (1992) considered massive black hole accretion and photodisintegration of He, but this is also ruled out. Cosmic strings have been ruled out as a primary source of the general anisotropy, and consequently, it is unlikely that there could be a direct effect on the CMB spectrum through energy release from the strings. On the other hand, a search for the spatial signature (anisotropy) of cosmic strings continues to be an active research area as better maps are obtained. Energy release by superconducting strings was considered by Ostriker and Thompson (1987) and Tashiro et al. (2012), and magnetic fields were discussed by Jedamzik et al. (2000).

2.11 Reionization History and the X-ray Background

The reionization of the intergalactic medium, presumably by UV from discrete sources such as the first stars and AGN, was accompanied by energy delivered to the CMB through the Compton scattering of the newly freed electrons. The reionization energy of about 10 eV per electron is small compared with the CMB energy of $2(1+z)$ MeV per electron, for $\Omega_B h^2 = 0.0125$ (Wright et al. 1994). The electrons left after decoupling, or produced by reionization, will be cooled fairly quickly by Compton processes until $z = 5$. If they are produced later, they might be relativistic and the CMB distortion need not follow a simple y form.

The hot electrons producing the X-ray background can be isotropically but not uniformly distributed; otherwise, they would produce a measurable distortion of the CMB spectrum. The FIRAS limit was that a hot IGM produces less than 10^{-4} of the X-ray background and conversely that the X-ray background comes from material with a filling factor less than 10^{-4}.

The Sunyaev–Zel'dovitch (Sunyaev and Zel'dovich 1970a) effect is the result of Compton scattering by hot gas in clusters of galaxies and has now been observed directly as hot spots in the sky maps (at short wavelengths) and cold spots at long wavelengths. With enough precision, it should be possible to observe the cumulative effect of these clusters on the mean sky spectrum and to determine whether the cluster population explains the whole distortion seen.

2.12 Far-IR Background Sources

It is expected that AGN and star-forming galaxies would have produced significant far-infrared fluxes from dust emission and that if this occurred at a high enough redshift, it might add to the CMB spectrum in a way that would be hard to recognize. In total the far-IR background measured by the COBE DIRBE and FIRAS instruments is comparable in brightness to all the known categories of visible and near-IR sources so that of order 1/3 of the total (post-recombination) luminosity of the universe is in the far-IR. The measured far IR background has been partially resolved into discrete sources, particularly ULIRGs at modest redshift ($z = 1 - 2$). This is an active field of study using data from the ISO, Spitzer, and Herschel space observatories, the BLAST instrument, far-IR cameras on large ground-based telescopes, and submillimeter interferometers.

3 High-Precision Spectrum Measurements

The history of the measurements of the CMB spectrum is full of challenges and mistakes, later overcome by clever design, diligent pursuit of systematic errors, and careful measurement of foreground sources of radiation that might be mistaken for CMB. When the CMB was first predicted in 1948, a measurement had already been made by Dicke's group at MIT, producing an upper limit of 20 K, but no connection to the prediction was made at the time. Wright (2012) gives an online tutorial and argues that Dicke might have been able to discover the CMB if he had tried. But by the time Penzias and Wilson were hunting for excess noise in their system, receiver sensitivity had advanced orders of magnitude, enough that the extra 3 K they saw was a significant part of the equipment sensitivity. In addition, they knew that the full sensitivity

Table 13-1

Summary of spectrum distortion parameter limits, with 95% confidence, or 2 σ. COBRA, FIRAS, and PIXIE report y, ARCADE 2 and TRIS report Y_{ff}. The proposed PIXIE mission would have sensitivity to measure many predicted spectrum distortions

Name	Freq. (GHz)	Y_{ff} or y	μ or $\lvert\mu\rvert$
COBRA (Gush et al. 1990)	90–480	0.002	0.008
FIRAS (Fixsen et al. 1996)	60–630	$-1 \pm 7 \times 10^{-6}$	$-1 \pm 4 \times 10^{-5}$
ARCADE 2 (Seiffert et al. 2011)	3–90	$<1 \times 10^{-4}$	$<6 \times 10^{-4}$
TRIS (Gervasi et al. 2008)	0.6–2.5	$[-6.3, -12.6] \times 10^{-6}$	$<6 \times 10^{-5}$
PIXIE (Kogut et al. 2011)	30–6,000	2×10^{-9}	10^{-8}

they wanted for telecommunication and astronomy would require an excellent antenna with very low sidelobes (response to signals from directions outside the main beam of the antenna). Also, scientific research projects now demanded liquid helium, which could be purchased commercially, so it was possible to design a calibration process that would compare the sky with a reference body at a low temperature. Finally, they were well aware that the Milky Way galaxy is bright and that the atmosphere might emit radiation, so they knew that they should measure these effects. All of these advances were crucial to their discovery and to the confirmation by the Princeton group months later.

In this section we first describe the foregrounds that obstruct precise measurements of the CMB spectrum and then describe the precise measurements already completed and the preparations now being made for even better measurements. Key projects include the COBE-FIRAS; its immediate confirmation by a similar but brief COBRA rocket experiment by H. Gush et al. and a balloon payload (ARCADE, Absolute Radiometer for Cosmology, Astrophysics, and Diffuse Emission) that has already flown to measure the cm-wave spectrum precisely; and the TRIS ground-based measurement at 0.6, 0.82, and 2.5 GHz. We discuss a proposed satellite mission (PIXIE, Primordial Inflation Explorer) that in principle could measure the CMB spectrum, anisotropy, and polarization and briefly describe the DARE (Dark Ages Radio Explorer), a proposed lunar satellite to search for redshifted 21 cm HI radiation that can be seen as spectral fluctuations of the CMB. We conclude with a discussion of the ultimate limits to CMB spectrum measurements.

We summarize the main results of the precise spectrum measurements in *Table 13-1*.

3.1 Minimizing Foregrounds

There are both diffuse and discrete foreground sources that limit the accuracy of CMB measurements. Excellent summaries have been published by both the WMAP (Gold et al. 2011; Bennett et al. 2003a) and Planck science teams (Planck Collaboration et al. 2011a and 25 additional articles in a dedicated journal issue), primarily addressing the effects on anisotropy measurements. Neither WMAP nor Planck measures absolute brightness, and hence both are insensitive to potential isotropic foreground radiation sources. Below, we discuss the Earth and its atmosphere, the Solar System, and Galactic electrons, dust, atoms, ions, and molecules.

3.1.1 Earth and Atmosphere

Terrestrial instrumentation must be protected from the brightness of the Earth, occupying 2π sr of solid angle immediately under the equipment. Typically the shield is made of reflective material but has a residual emissivity itself, which must be measured by for instance changing its temperature, varying the angles of incidence, etc.

In addition, the primary antenna must be made with extremely low sidelobes, and ideally those sidelobes must be mapped as a function of direction, polarization, and frequency. Many designs have been used. The Bell Labs antenna was called a Hogg horn, combining a very large pyramidal horn antenna with an off-axis parabolic reflector operating at 45° incidence to make a pattern with low response toward the ground. Other ground-based and low-altitude equipment has typically used much smaller horns with scrupulous attention to detailed design and test. One successful design uses a corrugated circular horn. These antennas have the advantage that the beam pattern is nearly Gaussian and nearly independent of polarization.

Above that is the atmosphere, which has a temperature gradient with altitude, moving invisible clouds of water vapor and visible clouds of water droplets or ice crystals. In mountain areas, the stratification of the atmosphere is not necessarily plane parallel, due to the wind blowing over the topography. The atmosphere also has complex chemistry, including the possible presence of water-vapor dimers that produce a continuum opacity. The primary atmospheric opacity at wavelengths where we wish to measure the CMB is due to water vapor, and there are also oxygen lines at 60 and 120 GHz and many submillimeter lines of ozone. There are excellent models for all of these except the water vapor dimers. Measuring and compensating for the atmospheric emission and opacity was critically important for ground-based measurements of the CMB spectrum. The primary technique is to measure the dependence of the measured CMB temperature on zenith angle.

3.1.2 Solar System

The diffuse zodiacal dust in the solar system, originating from the collisions of asteroids and comets with each other, is not bright at CMB wavelengths but can be detected. Its spatial distribution is concentrated in the ecliptic plane, but the cloud is thick and the contrast from ecliptic pole to ecliptic plane (at 90° elongation) is only about 1:3. According to theory and confirmed by measurements with the DIRBE instrument on the COBE mission, the dust density falls off with distance from the Sun as a power law $\rho \propto r^{-\alpha}$, where $\alpha = 1.34$ (Kelsall et al. 1998). The far-IR spectrum of the zodiacal dust was reported by Fixsen and Dwek (2002), based on the FIRAS and DIRBE observations. At wavelengths longer than 150 μm, the dust appears to have an emissivity proportional to ν^2, as expected from the Kramers-Kronig relations and causality, for dust grains significantly smaller than the wavelength, sufficiently hot, and not spinning.

In any case, little can be done currently to avoid this foreground emission, as it is difficult and expensive to put the observatory outside the zodiacal dust cloud. But in the future, missions to the outer solar system will become more feasible, propelled by ion engines and powered by more efficient radio-thermal generators, nuclear reactors, or even huge solar collectors.

3.1.3 Galactic Electrons

Galactic electrons collide with protons to produce free-free (bremsstrahlung) radiation and orbit in magnetic fields to produce synchrotron radiation. Both have spatial structure that is highly concentrated toward the Galactic plane and the Galactic center, and both follow approximate power-law spectra because the underlying electron energy spectra also have nearly power-law form. As shown in *Fig. 13-2*, below about 1 GHz, the synchrotron radiation is brighter than the CMB, and below about 0.5 GHz the free-free emission is also. We estimate the contribution of these sources to the CMB spectrum by mapping the entire sky at longer wavelengths, where the electrons are dominant, and predicting the maps at higher frequencies. There are subtleties related to deviations from simple power-law behavior, variations of the power-law indices, and the effects of discrete sources and polarization on the modeling. In addition, differences between the angular resolution and sidelobe response of the antennas used at different frequencies must be modeled. Considering all these complexities, precise experiments are designed to have a wide range of wavelengths with the same antenna pattern at all wavelengths. The WMAP team reported that the spectrum steepens between 20 and 40 GHz, consistent with steeper spectrum synchrotron sources, or with spinning dust like models (Gold et al. 2011). They conclude that masking out the bright regions of the Galaxy leaves negligible contribution to the CMB anisotropy or polarization. Similar methods could be used to subtract Galactic foregrounds from future high-precision CMB spectrum measurements like PIXIE.

Early results from the Planck mission confirm the WMAP result that the Galactic emission is not fully represented by the simple models of free-free, synchrotron, and dust emission. For example, there is a widely extended residual "haze" around the galactic center. Maps of the haze resemble the maps of gamma-ray bubbles observed with the Fermi mission. Hooper and Linden (2011) consider the possibility that dark matter annihilation produces the haze. In any case, it is likely that the electron spectrum does not have the same power-law index everywhere, and indeed nature does not require it to be a power law at all. This is an active research area for radio astronomers as well as cosmologists.

3.1.4 Galactic Dust

As shown in *Fig. 13-2*, at wavelengths less than about 0.5 mm ($\nu > 600$ GHz), the thermal emission from interstellar dust grains is brighter than the CMB. Although interstellar dust grains cannot be easily collected, evidence suggests that there are many different kinds, shapes, sizes, and compositions and that they are not all at the same temperature even within a single cloud of dust and gas. Their properties evolve with time as they are sputtered by cosmic rays, or are heated in shocks or grain-grain collisions or near approaches to hot stars, or serve as condensation nuclei in cold gas clouds. Moreover, the grains can be aligned by magnetic fields, as we know from the polarization of starlight and scattered starlight, and therefore can emit partially polarized far-infrared radiation. In addition, the smallest grains can spin at frequencies of 10–100 GHz. They rotate because they absorb and scatter light preferentially from locations away from the center of mass, and their rotation rates are limited by their mechanical strength. They are also struck occasionally by high-energy atoms and cosmic rays (Draine and Lazarian 1998, 1999).

The nonspinning dust emission is typically represented by a power-law emissivity modifying a Planck function: $I_\nu \propto \nu^\beta B_\nu(T)$, with $\beta \approx 1.8$, not far from the Kramers-Kronig limit of 2. However, this effective index may simply approximate the sum of dust populations at different temperatures. The fitted temperatures range from 10 to over 20 K, depending on direction, and sometimes more than one component is required for a single direction.

3.1.5 Galactic Atoms, Ions, and Molecules

Only a few species of Galactic atoms, ions, and molecules emit strongly enough to be detectable (so far) in the wide bandwidths needed to measure the spectrum of the CMB. These are CO (up to the 8–7 line), [C I] at 370 and 690 μm, [C II] at 158 μm, [N II] at 122 and 205 μm, [O I] at 146 μm, and CH at 116 μm (Fixsen et al. 1999). They also found the 269 μm line of H_2O in absorption against the Galactic center. The CO lines are dominant cooling lines for cold clouds, and the [C II] line emits about 0.3% (Wright et al. 1991) of the dust luminosity of the Milky Way, so they are of great interest for astrophysics. Because these lines are concentrated in the Galactic plane and in molecular clouds, they have not limited the accuracy of CMB spectrum measurements. On the other hand, the CO lines appear in the passbands of some instruments designed to measure CMB anisotropy.

The lines also provide frequency calibration standards for CMB spectrometers. From these calibrations and the known values of the Planck and Boltzmann constants, it was possible to confirm and improve the thermometric accuracy of the COBE FIRAS.

3.2 COBE FIRAS

The most precise (50 parts per million) measurement of the CMB spectrum yet accomplished was made with the FIRAS (Far Infrared Absolute Spectrophotometer) on the COBE satellite (Fixsen et al. 1996). The COBE (Boggess et al. 1992) was launched Nov. 18, 1989 from Vandenberg Air Force Base near Lompoc, California. It still orbits 900 km above the Earth in a polar (94° inclination) Sun-synchronous orbit so that the orbit plane precesses at 1 rotation/year to remain approximately perpendicular to the line to the Sun. The COBE carries three instruments, all protected by a conical shield. The COBE was oriented so that the Sun was always a few degrees below the plane of the top of the shield, so the instruments were well protected. However, for about 2 months per year, the Earth limb rose slightly above the shield plane for about 20 min out of each 103 min orbit. A higher altitude orbit could have avoided this situation but would have caused higher rates of Van Allen belt particle bombardment. The FIRAS and DIRBE (Diffuse InfraRed Background Experiment) instruments were cooled to about 1.5 K inside a liquid helium cryostat. The third instrument, DMR (Differential Microwave Radiometers), was mounted around the circumference of the cryostat and inside the shield.

3.2.1 FIRAS Design

The FIRAS was sensitive to wavelengths from 100 μm to 1 cm, including the peak emission of the CMB around 2 mm, the Wien tail of the distribution for $\lambda < 1$ mm, and strong emission from the interstellar dust. The instrument was based on the design of the balloon-borne spectrometer flown by UC Berkeley in the mid-1970s (Woody et al. 1975; Woody and Richards 1979). The instrument is a rapid-scan, symmetrical polarizing Fourier transform spectrometer with two inputs and two outputs, using the Martin and Puplett (1970) concept. The 7° instrument beam is defined by a Winston cone, a quasi-optical compound parabolic concentrator (Winston 1970). The orbit and orientation of the spacecraft meant that the spin axis and FIRAS line of sight were approximately 94° from the Sun at all times and so swept out an approximately great circle as the satellite orbited the Earth. The circle moved with the Sun so that over the course of 6 months, the beam scanned the entire sky, and in the 10 month duration of the liquid helium, about 60% of the sky was observed at intervals 6 months apart. The Winston cone attaches smoothly to an apodizing section, flared like a trumpet bell. Calibration was provided by a full-beam external blackbody that could be moved into the antenna, where it sealed the aperture. The interferometer mirrors were scanned smoothly in a sawtooth pattern, producing a time-dependent intensity at each detector, and there were two stroke lengths and two stroke speeds that could be chosen by command. The interferometer functions as a modulator that compares two inputs, one from the sky (or external calibrator) and one from an internal reference body. Each output of the instrument is split by a dichroic beamsplitter into short and long wavelength bands, separated at $\lambda = 0.5$ mm, so that there are four detectors altogether. Combining the two scan speeds, the two stroke lengths, the two sides, and the two wavelength bands, there are altogether 16 independent data sets (➋ *Figs. 13-3* and ➋ *13-4*).

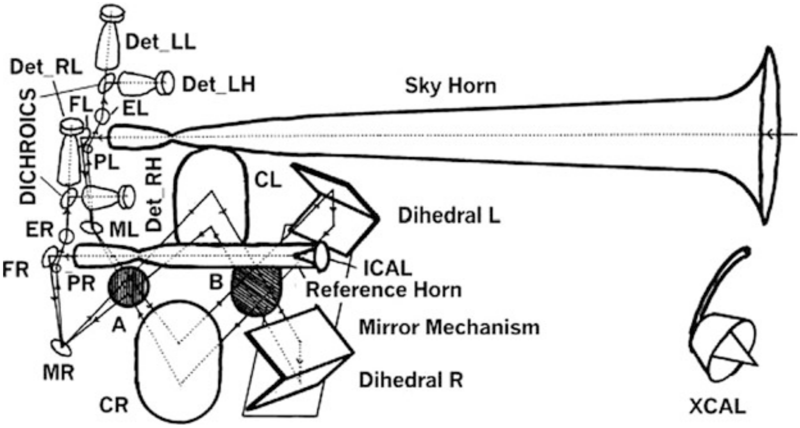


Fig. 13-3
Drawing of the FIRAS instrument. Light enters the sky horn from the sky or the XCAL and the reference horn from the ICAL. After reflection from the folding flats (FL, FR), it bounces off the mirrors (ML, MR) and is analyzed by the polarizer (A). The collimator mirrors (CL, CR) recollimate the light before it is split by a second polarizer (B) at 45°. It is then reflected by the dihedral mirrors, with different paths set by the mirror mechanism. After reflection, the light retraverses the beam splitter, collimator mirrors, and analyzer. This time it is intercepted by the pickoff mirrors (PL, PR), which direct it into the elliptical mirrors (EL, ER), the dichroic filters, and, finally, the detectors (DetLH, DetLL, DetRH, DetRL)

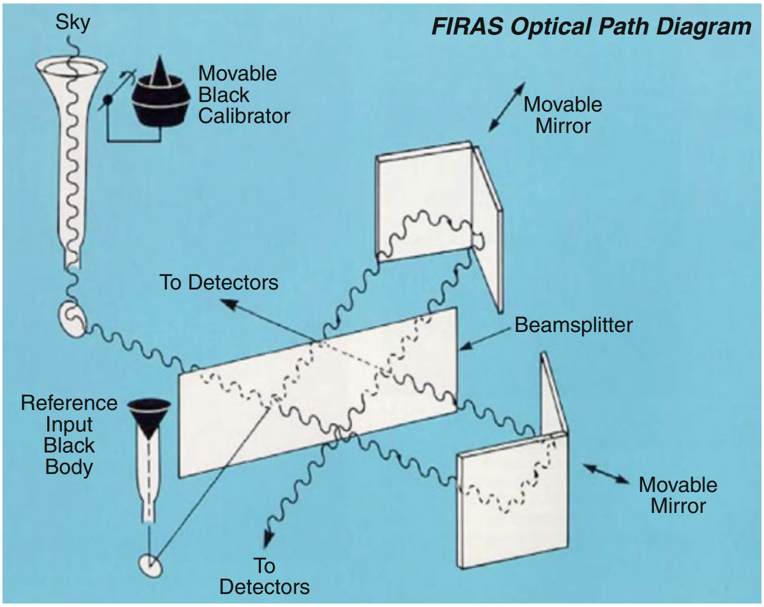


Fig. 13-4
FIRAS calibration concept. The actual FIRAS used two wire grid polarizers with wires at 45°. The external calibrator nearly sealed the horn antenna when in place. The internal reference body was adjusted to nearly null the sky interferogram. Both horns and both blackbodies were controllable from about 2 to 20 K

3.2.2 FIRAS Detectors and Data Processing

The four detectors were composite bolometers, with noise equivalent powers of the order of NEP $= 4 \times 10^{-15}$ W/Hz$^{1/2}$. Each used a doped silicon thermometer chip, bonded to a large but thin diamond octagon, blackened with a thin film of bismuth, and all suspended by taut Kevlar fibers. Each was DC biased through a large resistor, and the signal was amplified by a heated JFET to feed the signal out the long coax cable to the warm electronics. The JFET was suspended inside a radiation-tight box on Kevlar fibers so that it could operate near its optimum temperature of about 60 K. The sensitivity was limited by charged particle impacts, which were very common in certain parts of the orbit near the horns of the Van Allen belts around the Earth's poles, and in the South Atlantic Anomaly, where the radiation belts are especially near to the Earth.

Data processing began with sorting the interferograms into groups according to observing mode, temperature settings of the horns and calibrators, detector bias, and line of sight. Then, interferograms were compared to optimally detect the impulses from cosmic rays, and these signatures were iteratively removed by a least squares fitting program.

3.2.3 FIRAS Calibration

The operation of FIRAS in the vacuum and cold of outer space enables a nearly ideal configuration for comparison of the sky with a full-beam blackbody. If the calibrator body is perfectly black and isothermal, and there is no change of the instrument signal when the body is inserted or removed from the beam, then the sky also has a perfect blackbody spectrum, regardless of the imperfections or calibration factors of the instrument or the uncertainty of the thermometer calibration. To reduce the dynamic range of the instrument, the internal reference body was adjusted to nearly null the modulated interferogram by setting its temperature close to the sky temperature. To be more quantitative, temperature controllers were provided to regulate the temperatures of the external calibrator and the internal reference body and of the Winston cone (sky horn) and its counterpart facing the internal reference body.

The calibration (Fixsen et al. 1996) was designed to measure the parameters of an instrument model that included the emissivities of the Winston cone (sky horn), the similar but smaller horn (reference horn) receiving signals from the internal reference body, the emissivity of the reference body, and systematic and temporary errors in thermometry. Although all the thermometers were calibrated to mK accuracy before installation, there was evidence that the calibrations were somewhat incorrect in flight. Fortunately, the thermometer errors do not limit the determination of whether the sky has a blackbody spectrum, but they do limit the accuracy of the determination of the absolute temperature. There were also small residual errors in the calibration data sets that showed that the mirror transport mechanism (MTM) was vibrating as it moved, as was known before launch, and that some light was making more than one pass through the interferometer, yielding the appearance of harmonic response. Models were made for these effects, and their parameters were determined from the data. In the end, there was a single calibration model representing all the observing modes and all the detectors, with a full covariance matrix for the errors (➋ *Fig. 13-5*).

The accuracy of the calibrator body is critical, and the design is reported by Mather et al. (1999). It is made of Eccosorb CR-110, cast in the form of a reentrant cone, like a trumpet

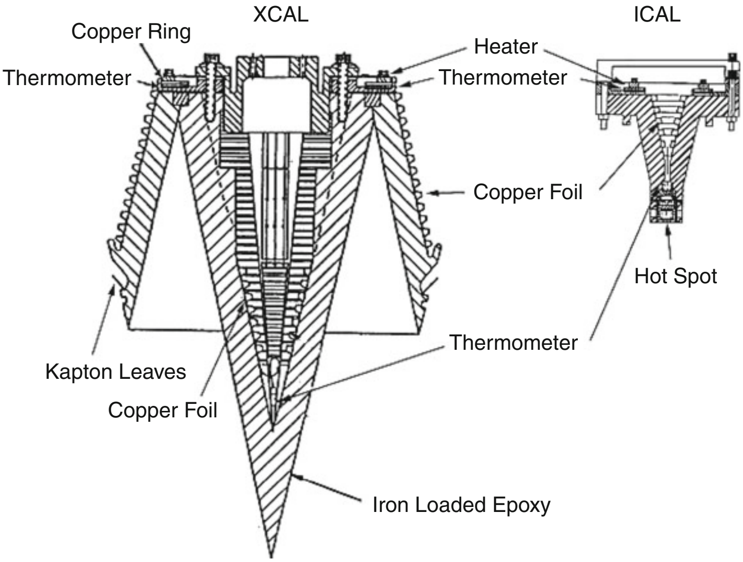


Fig. 13-5
Cross section of the FIRAS calibrators. The XCAL is 140 mm in diameter, and the ICAL is 60 mm in diameter. Heaters and thermometers are indicated on the drawing. The Hot Spot heater was designed to null a high-frequency excess in the CMBR. No excess was seen, but the Hot Spot is part of the reason the ICAL has a reflectance of ~4%

mute. This is an epoxy loaded with iron powder to increase its permeability and absorption of microwaves; in this case it was also mixed with a fine silica powder to make it thixotropic (to resist the settling of the iron powder during casting). The calibrator material is cast onto a corrugated copper foil to control its temperature and make it as isothermal as possible. The support arm and the top of the calibrator are wrapped with multilayer insulation to protect against radiant heat flow from the warm rim of the sunshield.

There are several ways in which the calibrator could be imperfect. First, it could be partially reflective so that emission (or lack thereof) from parts of the spectrometer that are not at the same temperature would change the effective input. Since the spectrometer includes parts at many temperatures, this was the main worry. The relevant parts include the detectors, the mirrors, the beamsplitters, the horns, and the internal reference body, as well as the surrounding chamber. Only the horns and the internal reference body could be actively controlled. The effective reflectance of the calibrator in the horn was made as low as possible and was estimated to be less than 10^{-5}, depending on frequency. It was also measured with a value of -55.8 ± 1.5 dB at 33.4 GHz. When the calibrator is in the FIRAS horn, any ray reflected from the calibrator will bounce from the horn surface many times and almost always return to the calibrator for another chance to be absorbed, so the effective reflectivity of the calibrator in the horn is much lower than the actual reflectivity. (The other absorbers in this cavity are the aluminum horn and the exit port to the spectrometer.) Other potential errors include partial transparency of the calibrator material, allowing leakage from the insulated outer surface, which was slightly illuminated at some angles by rays from the solar shield around the cryostat. This effect, if important, would be maximum at the longest wavelengths, where the calibrator material had the least absorptivity, but no sign of a problem was seen. A third potential effect would be leakage through the gap between the calibrator and the horn. To control this leakage, two sets of aluminized Kapton leaflets were installed around the circumference of the calibrator, to make a direct optical seal of the gap, without transmitting any heat in case the calibrator and horn were at different temperatures. Measurements were made of the calibration signal as a function of the position of the calibrator in the horn, and no effect was seen until the calibrator had been moved about 12 mm. A fourth effect would be if the calibrator were not isothermal. In this case, it would emit a spectrum like the cosmic y distortion and enable a real cosmic y distortion to escape notice. However, the measured temperature gradients, and the known thermal conductivity, together with models of the potential heat flows, show this to be negligible on the scale of the FIRAS sensitivity.

Fixsen et al. (1996) and Fixsen (2009) developed alternate ways to redetermine the temperature scale. The wavelength scale of the FIRAS was determined precisely using the atomic and molecular spectrum lines of the Galaxy and used to cross-check the observations of the temperature dependence of the calibration signals. Also, the spectrum of the dipole anisotropy of the CMB as measured by FIRAS has its own spectral form (the derivative of the Planck function with respect to temperature) and can be used to determine yet another estimate of the monopole temperature. Fortunately, all these measurements are consistent at the level of 1 mK. Finally, the FIRAS results can be recalibrated using the WMAP dipole.

3.2.4 Foreground Removal for FIRAS

At the wavelengths observed by FIRAS, the main foreground is the dust emission from the Galaxy, concentrated in the Galactic plane. An adequate (in the sense of good χ^2) model for the dust emission in most lines of sight is a product of a power-law emissivity as a function of frequency and a Planck function for a single temperature. However, in the Galactic plane where the dust is bright, there were many regions where this was not an adequate fit, and two or even three components at different temperatures were required. At high Galactic latitudes, the dust is too faint to measure in most individual lines of sight, so a smoothed Galactic model was used to fit the data. The smoothed model was not quite proportional to the cosecant of the Galactic latitude.

The spectrum lines of the interstellar atoms, ions, and molecules had a negligible effect on the measurement of the CMB spectrum because they are either faint, concentrated in the Galactic plane and Galactic center, or at frequencies well above the CMB frequencies. The [C II] line at 158 μm is so bright that it produces a recognizable cosine curve in the raw interferograms of the Galactic plane.

3.2.5 Limits on CMB Spectrum Distortion

Least squares fits to the FIRAS data produced the limits in Fixsen et al. (1996). The fitted parameters are correlated because the y and μ distortion spectra are not orthogonal to the derivative of the Planck function with respect to temperature $\partial B/\partial T$ or to the model spectra for the interstellar dust. In addition, the calibration model parameters are themselves correlated. The results were $|\mu| < 10^{-5}$ and $|y| < 15 \times 10^{-6}$, 95% confidence, including systematic errors. The RMS deviation between the measured spectrum and the blackbody is less than 50 parts per million of the peak brightness.

The most precise measurement of the CMB temperature was reported by Fixsen (2009), giving the value $T_{\text{cmb}} = 2.72548 \pm 0.00057$ K, based on recalibration of the FIRAS data in comparison with the WMAP data, and other published data. This precise value is important for comparison with measurements at different wavelengths.

Fixsen et al. (1997) showed that the spatial fluctuations of the CMB as observed by the FIRAS instrument on a 7° scale have a thermal spectrum (actually, the derivative of the Planck function with respect to temperature), as they should if they are really temperature fluctuations. Stated another way, the FIRAS, an absolute instrument, detects the CMB anisotropy. The data also limit rms fluctuations in the Compton y parameter, observable via the Sunyaev–Zel'dovich effect, to $\Delta y < 3 \times 10^{-6}$ (95% CL) on 7° angular scales.

3.3 COBRA Rocket Experiment

The COBRA rocket instrument of Gush et al. (1990) was launched for the fifth time on Jan. 20, 1990, just 2 months after the launch of COBE. Built at the University of British Columbia, it included a rapid-scan Fourier transform spectrometer very similar in concept to the COBE FIRAS instrument, but using a dielectric beamsplitter instead of the polarizing configuration. In its 9 min of flight, it obtained excellent quality data, using a much colder detector (0.29 K) than the FIRAS to gain far higher sensitivity. In its short flight, there was no time or space to move an external calibrator body into the beam, so the instrument was calibrated on the ground before launch, using an external calibrator with emissivity >0.999. There was also no time to scan the sky or make a map. Nevertheless, it was immensely important that this instrument also observed the blackbody and that its measured temperature (2.736 ± 0.017 K) agreed well with the FIRAS result. (There was still the possibility that the FIRAS thermometers were not correctly calibrated.) The reported 95% confidence limits on y and μ were 0.001 and 0.008, respectively (➋ *Fig. 13-6*).

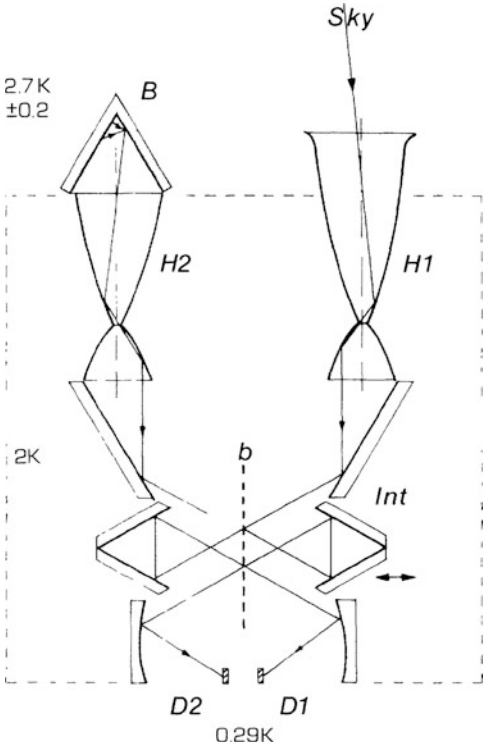


Fig. 13-6
COBRA design of Gush et al. (1990) showing the principle of the apparatus. H1 and H2 are similar horn-type telescopes of 6° field of view. H1 receives radiation from the sky, whereas H2 is illuminated by a blackbody simulator B. Radiation issuing from the horns enters a two-beam interferometer (Int), with a beam splitter at b, from whence it emerges to be focused on two bolometric detectors, D1 and D2. As the path difference in the interferometer is changed, the signal generated by each detector (interferogram) is proportional to the *difference* in intensity of the sky and the blackbody. The numbers indicate temperatures of various sections of the spectrometer. One challenge of the design was making a reference load whose temperature could be changed reliably during the brief rocket flight.

3.4 ARCADE 2

The ARCADE 2 instrument (Absolute Radiometer for Cosmology, Astrophysics, and Diffuse Emission, described by Singal et al. (2011)) is the most ambitious yet built for measurement of the CMB spectrum at wavelengths out to 10 cm (3 GHz). It is a balloon-borne microwave radiometer with six frequencies (3, 5, 8, 10, 30, and 90 GHz), a double-nulled design with internal reference bodies and an external calibrator, and is completely windowless to enable all the radiometrically active parts to be cooled to 2.7 K. It extends the FIRAS concept as far as possible to long wavelengths, given the size and atmospheric constraints of balloon payloads. To improve on ARCADE 2 will require a redesigned top surface area or a space mission.

The key discovery that enables this design is that it is possible to achieve adequate flow velocities for gaseous helium, escaping in a controlled way from a 1,900 liter cryostat, to keep the

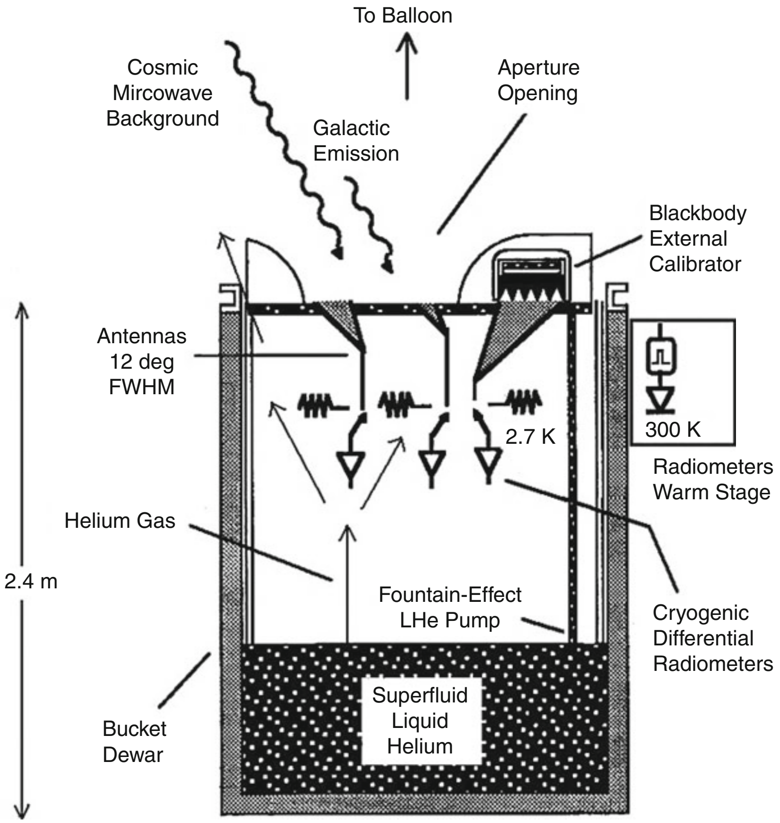


Fig. 13-7

ARCADE 2 instrument schematic, components not to scale, as described by Singal et al. (2011). Cryogenic radiometers compare the sky to an external blackbody calibrator. The antennas and external calibrator are maintained near 2.7 K at the mouth of an open bucket dewar; there are no windows or other warm objects between the antenna and the sky. Cold temperatures are maintained at the *top* of the dewar via boil-off helium gas and tanks filled with liquid helium fed by superfluid pumps in the bath. For observing the sky, everything shown is suspended below a high-altitude balloon

atmosphere from condensing on the apparatus. Scale models of the aperture were built to simulate the flow on the ground, and in-flight cameras confirmed success. In addition, superfluid fountain-effect pumps lifted up to 55 l/min of superfluid helium to the top of the apparatus, to keep it cold. This is a remarkable feature of engineering that enables nearly ideal calibration to be performed. Although the calibrators are in flowing gaseous helium, they can reach the desired temperature of 2.725 K. The details and variations of heat flow and temperature gradients in the calibrator bodies are then the limiting factors for calibration accuracy (*Fig. 13-7*).

The ARCADE 2 used corrugated circular antenna feeds to produce 11.6° FWHM Gaussian beams. In addition, it carried a 4° beam antenna at 30 GHz for useful cross-checks. All observed at 30° from the zenith, away from the balloon and the payload suspension (parachute, ladder, truck plate, and FAA transmitter). The beam profile of one of the horns was measured and agreed extremely well with the theoretical predictions; it is almost independent of frequency within each observing band. The calculated radiometric effects of the balloon and suspension ranged from 2.3 to 42 mK, depending on frequency band.

The radiometers all used cryogenic HEMT amplifiers, which set the system sensitivity, followed by coax or waveguide links to warm amplifiers. The input to each cryogenic HEMT was switched to alternate between sky and reference load using a MEMS switch at 3 and 5 GHz and a latching ferrite waveguide switch for the higher frequency channels. The cold loads were in coax at the 3 and 5 GHz bands, and were made of steelcast wedges in waveguide for the other channels. The warm amplifiers were followed by frequency filters that divided each band into two parts, low and high, for improved spectral information.

3.4.1 ARCADE 2 Calibration

The critically important external calibrator was moved from one receiver aperture to another, so all shared the same thermometry and almost the same thermal environment, just as for the COBE-FIRAS. A precursor calibrator is described by Kogut et al. (2004) and the actual device by Fixsen et al. (2006). The calibrator surface was made from 298 sharp cones of steelcast, each 88 mm long and 35 mm in diameter, cast onto an aluminum core for thermal control. The measured reflectance of the calibrator ranged from −42.4 dB at 3 GHz to −62.7 dB at 10 GHz. These reflectances are so low that they are not a significant source of error. The harder problem is thermal gradient control, considering the flowing gaseous helium and the superfluid helium pumps. The measured thermal gradient from tip to base of the cones is 600 mK, but 98% of it occurs near the tip and involves only 3% of the absorber. The calculated calibrator emission was based on the measured thermal gradients, the known absorption properties of the material, and the known field distribution of the antenna pattern. There are systematic differences in heat flow according to the position of the calibrator over the different receiver apertures, but these were measured with many sensors. The calibration accuracy of ARCADE 2 could be improved up to an order of magnitude, by providing active control of the temperature of the top plate, to near 2.7 K instead of the 1.4 K achieved in the last flight.

3.4.2 ARCADE 2 Results

Seiffert et al. (2011) report the final results from the ARCADE 2 instrument after accounting for all calibration issues and the Galactic foreground. They find 2σ limits on the spectral distortion of $\mu < 6 \times 10^{-4}$ and $|Y_{\mathrm{ff}}| < 1 \times 10^{-4}$. They also find that there is a residual signal that significantly exceeds the models of all known galactic and extragalactic sources. It has a power-law spectrum with amplitude 18.4 ± 2.1 K at 0.31 GHz and a spectral index of −2.57 ± 0.05 (➲ *Fig. 13-8*).

Vernstrom et al. (2011) consider the extragalactic source population in detail and conclude that their model does not explain the ARCADE 2 results. Seiffert et al. (2011) had suggested that there might be a population of sub-μJy sources down to 10 nJy at 1.4 GHz that add up to enough, and it might be possible if high z star-forming galaxies have a higher radio-to-far-IR ratio than local ones. Radio observations at greater depth than yet obtained could be made with the Expanded Very Large Array (EVLA) and the ALMA.

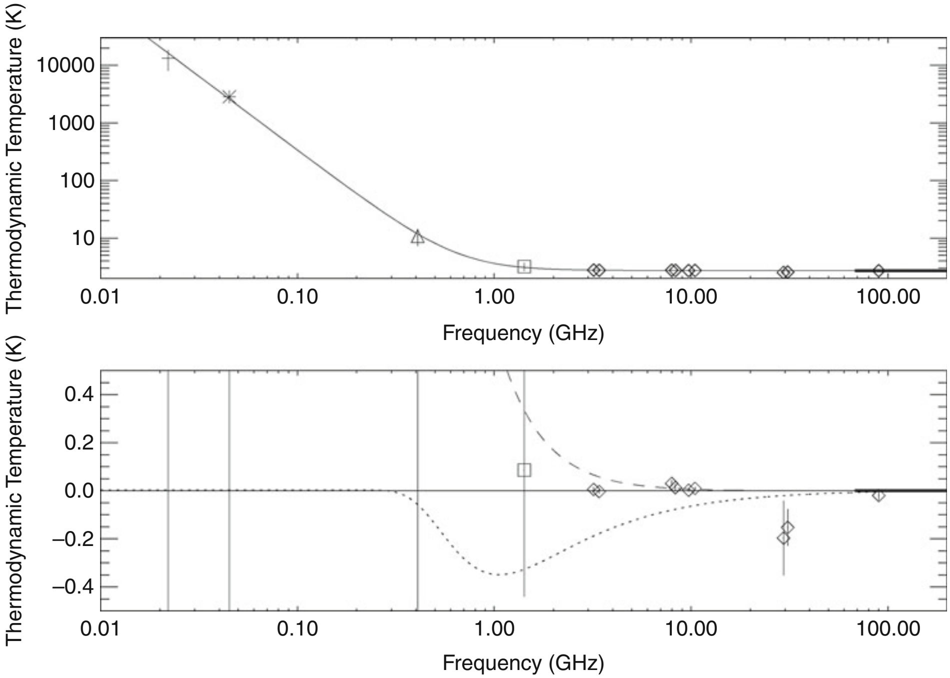


Fig. 13-8

ARCADE 2 results as described by Seiffert et al. (2011). Fit of ARCADE 2 data, FIRAS data, and data from low-frequency radio surveys. The *upper plot* shows (*solid line*) a fit with three components: a frequency- independent CMB contribution, a power-law amplitude, and a power-law index. The *lower plot* shows the fit residuals. The *dotted line* shows the expected shape of a μ distortion. The amplitude of the plotted distortion is 50 times the upper limit determined from FIRAS. The *dashed line* shows the shape of a Y_{ff} distortion with an amplitude equal to the 2σ upper limit. The addition of either a μ distortion or a Y_{ff} distortion as a free parameter is not supported by the data. Data points are from Roger et al. (1999), cross; Maeda et al. (1999), *asterisk*; Haslam et al. (1981), *triangle*; Reich and Reich (1986), *square*; ARCADE 2 (*diamonds*); and FIRAS (*heavy line*), corrected for Galactic emission and an estimate of extragalactic radio sources, as shown in Table 1 of Seiffert et al. (2011)

3.5 TRIS

Gervasi et al. (2008) report on the ground-based TRIS experiment (Zannoni et al. 2008), which improved the uncertainty of measurements at frequencies of 0.60 and 0.82 GHz by factors of 9 and 7, respectively, and agreed with prior measurements at 2.5 GHz. The improvements were obtained through better absolute calibration and better modeling of the Galactic foregrounds. The TRIS operated primarily by doing drift scans while pointed at the zenith. Absolute calibration was obtained using a cryogenic front end of each receiver, with a triple-throw switch to connect the receiver to the sky, a cold load (CL) near 4 K, or a warm load (WL) near 270 K (*Fig. 13-9*).

The results for the spectrum distortion parameters were $-6.3 \times 10^{-6} < Y_{\mathrm{ff}} < 12.6 \times 10^{-6}$ and $|\mu| < 6 \times 10^{-5}$, both at the 95% confidence level.

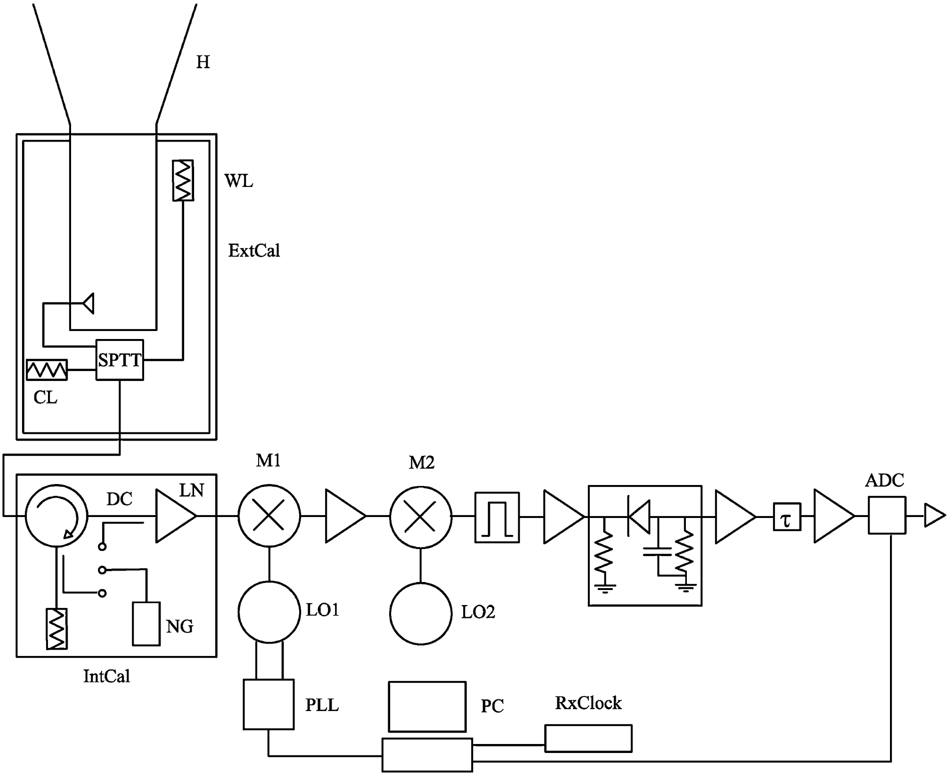


Fig. 13-9
Block diagram of TRIS radiometers. *H* corrugated horn, *LN* low-noise amplifier, *LO1, LO2* local oscillators, *M1, M2* mixers, *PLL* phase-locked loop, τ system time constant; *ADC* analog-to-digital converter, *PC* personal computer, *RxClock* radio clock, *ExtCal* external calibrator (*WL* warm load, *CL* cold load, *SPTT* switch), *IntCal* internal calibrator (*C* circulator, *DC* directional coupler, *NG* noise generator)

3.6 PIXIE

The proposed Primordial Inflation Explorer (PIXIE) is a nulling polarimeter for cosmic microwave background observations as described by Kogut et al. (2011). It improves on the FIRAS in several ways, and it is able to measure the polarization and anisotropy of the CMB as well as its spectrum. It is fully symmetrical, unlike the FIRAS, and either of the two inputs can observe the sky or a full-beam reference blackbody. It would cover the range from 30 GHz to 6 THz (1 cm–50 μm wavelength) in 400 spectral channels and could map the Stokes I, Q, and U parameters over the whole sky. While the primary objective in today's context is the search for the polarization signature of primordial gravitational waves, the ability to test the CMB monopole spectrum against a blackbody is remarkable (➋ *Fig. 13-10*).

Improvements over the FIRAS concept include the use of 550 mm-diameter primary reflectors to define the beams on the sky, instead of the quasi-optical Winston cone used by the FIRAS. The PIXIE design preserves the polarization sensitivity, which was intentionally ignored in the

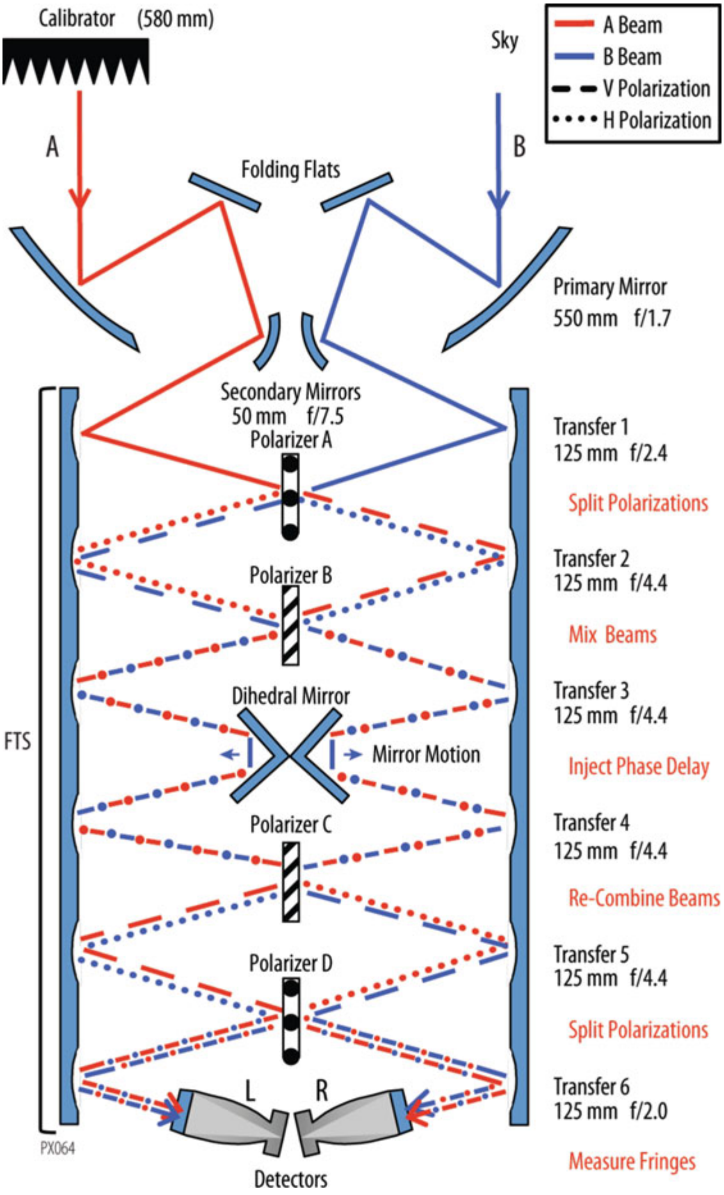


Fig. 13-10
Schematic view of the PIXIE optical signal path. As the dihedral mirrors move, the detectors measure a fringe pattern proportional to the Fourier transform of the difference spectrum between orthogonal polarization states from the two input beams (Stokes Q in instrument coordinates). A full-aperture blackbody calibrator can move to block either input beam or be stowed to allow both beams to view the same patch of sky

FIRAS. Each output is split into two polarization states following a square concentrator in front of the detectors so that there are four detectors in total. The beam size is smaller (2.6° top hat) than the FIRAS (7°) and is optimized for the low-order peak in the primeval B-mode polarization spectrum. Other instruments, many on the ground or balloons, appear capable of detecting the higher order polarization peak. But a full characterization of the low-order ($\ell < 20$) B-mode peak requires full sky coverage and a space mission. The étendue ($A\Omega = 4\,\mathrm{cm}^2\,\mathrm{sr}$) is much larger than for FIRAS ($1.5\,\mathrm{cm}^2\,\mathrm{sr}$), enabling good performance at longer wavelengths. The individual detector sensitivity is almost two orders of magnitude better (7×10^{-17} vs. $4 \times 10^{-15}\,\mathrm{W\,Hz}^{-1/2}$), based on the use of lower detector temperature (0.1 K instead of 1.4 K). In addition, the entire instrument (except the detectors) is cooled to the mean temperature of the CMB (2.725 K) so that emissions and reflections within the instrument do not produce modulated signals. Finally, the PIXIE would be cooled by active refrigerators (a mechanical cryo-cooler and staged adiabatic demagnetization refrigerators), as well as passive radiators, so that its much-longer (4 year) lifetime is not limited by stored cryogens (➔ *Fig. 13-11*).

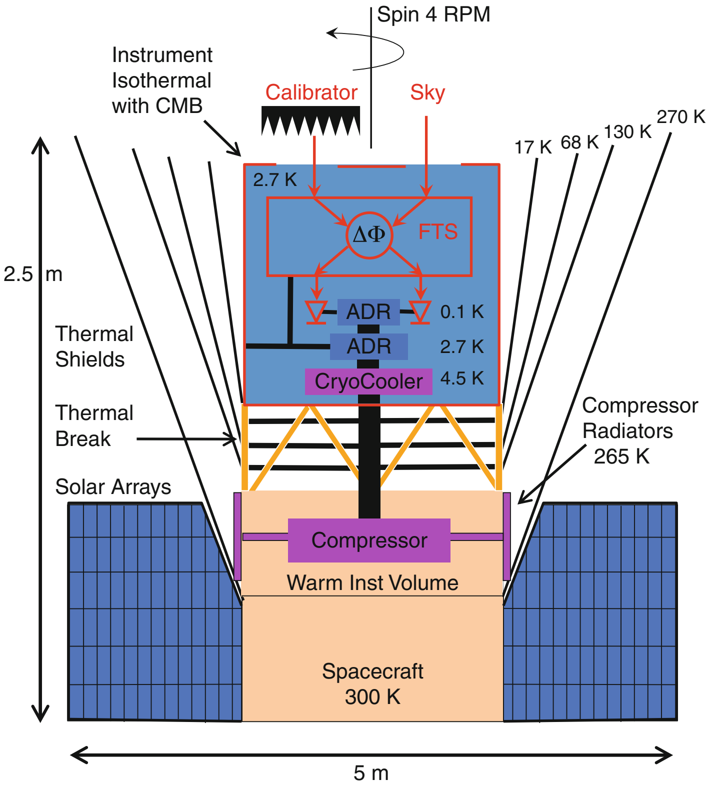


Fig. 13-11

Cryogenic layout for the PIXIE instrument. An ADR and mechanical cryo-cooler maintain the instrument and enclosure at 2.725 K, isothermal with the CMB. A set of concentric shields surrounds the instrument to prevent heating by the Sun or Earth. Approximate dimensions of the deployed observatory are indicated

The entire PIXIE observatory would spin at 4 rpm, much faster than the COBE (0.8 rpm), for rapid modulation of the polarization in the shorter time that it takes for the line of sight to move one beamwidth across the sky. The orbit altitude (660 km) is lower than for COBE (900 km), reducing the cosmic ray rates on the detectors.

The predicted PIXIE sensitivity for polarization is 70 nK per 1° square pixel, corresponding to multipole orders up to about 200. Averaged over the best 75% of the sky, it would have a sensitivity of 3 nK for the B-mode polarization signal, well under the 30 nK predicted from large-field inflation models, and comparable to the predicted noise floor from gravitational lensing of the CMB. For the spectrum distortions, after accounting for foreground removal, the predicted PIXIE sensitivity is μ of 10^{-8} or y of 2×10^{-9}. For comparison, the FIRAS limits were $|\mu| < 10^{-5}$ and $|y| < 15 \times 10^{-6}$, 95% confidence, including systematic errors. So the PIXIE would have a distortion sensitivity of three to four orders of magnitude better, opening up many possibilities for tests or detections of unexpected cosmic phenomena.

3.7 DARE

The Dark Ages Radio Explorer (DARE, Burns et al. 2012) is a proposed space mission to measure the very long wavelength spectrum of the CMB with enough precision to detect the variations due to the redshifted 21 cm hydrogen line. Since the galactic synchrotron and free-free emissions all have smooth power-law spectra, deviations from those forms would be meaningful. Operating in the 40–120 MHz range, the mission would be sensitive to hydrogen at redshifts from 11 to 35. To obtain a quiet-enough environment, the instrument would orbit the Moon and would observe when it is protected from both the Sun and the Earth. The predicted spectrum distortions range from +30 mK around 100 MHz from the Hot-Bubble-dominated epoch, to −110 mK around 65 MHz due to accreting black holes. Instrument calibration at these low frequencies is done with noise diodes, and the antennas are short dipoles and hence very nondirectional. The predicted spectral features are relatively narrow (10%) so that they would be recognizable even in the presence of strong Galactic emission.

3.8 Ultimate Limits

The PIXIE instrument pushes the limits of what will be possible, as its estimated sensitivity is 3–4 orders of magnitude better than the COBE FIRAS. Measurements to detect the spectral distortions from the individual recombination lines at decoupling would require parts per billion sensitivity, but on the other hand, a template for all of them together might be more easily recognizable. It might well be possible to achieve adequate raw sensitivity with an advanced mission, but at this level every foreground component will be bright, including molecules with narrow spectral features. Is the Planck function correct? The calibration data for the FIRAS instrument and the proposed PIXIE instrument could be used to test it very well. Note that the cosmological results do not require a precise verification of the Planck formula; all they require is that the sky and the blackbody calibrator must match.

4 Historical Overview of Temperature and Polarization Anisotropy

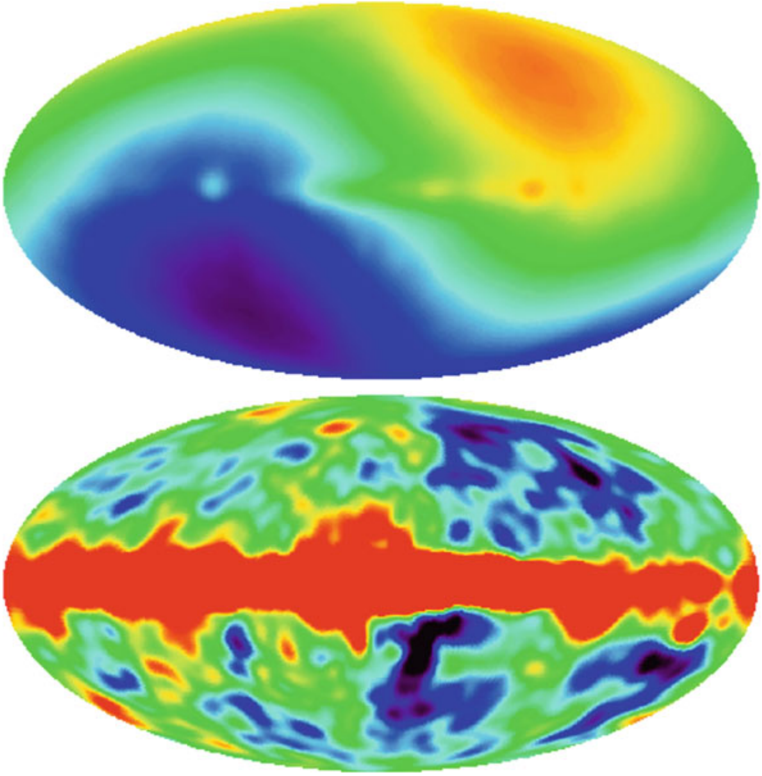
The cosmic microwave background (CMB) radiation was detected by Penzias and Wilson (1965). After its discovery, a small number of experimentalists worked for years to better characterize the CMB as they searched for temperature fluctuations. It is uncommon that the beginning, emergence, and maturation of a scientific field can be seen so clearly. The anisotropy was discovered nearly 20 years ago (Smoot et al. 1992). In that time, the temperature power spectrum has been mapped out through eight acoustic peaks, the best-fit cosmological model identified, and the polarization anisotropy (an independent check of the model, in this case) measured to high accuracy. As of this writing, we await the release of the Planck data. In the following, we briefly review the set of measurements that got us to where we are, summarize the current state of the data and its interpretation, and then indicate some future promising directions. In a separate contribution, Hanany et al. (2012), CMB instrumentation is discussed, so that is not included here.

4.1 Anisotropy Searches Prior to COBE

One of the primary reasons for believing that the Penzias and Wilson discovery was cosmological in origin was that the radiation was isotropic. Immediately, one wanted to know "how isotropic is it?" The instrumentation for measuring anisotropy is quite different than that used for the absolute temperature measurements. The first dedicated instrument was the Wilkinson and Partridge (1967) "isotropometer." The instrument had a beam width of 5.5° at a frequency of 9.4 GHz and was used to set an upper limit of 3.2 mK on the dipole amplitude at a declination of -8°, and a limit on smaller-scale anisotropy of $\sqrt{\ell(\ell+1)C_\ell/2\pi} < 2{,}700\ \mu\text{K}$ at $\ell \sim 10$. Given estimates of galaxy peculiar velocities, it was apparent early on (Peebles and Wilkinson 1968; Bracewell and Conklin 1968) that a dipole term should exist at an amplitude of roughly 3 mK. This term tells us our velocity with respect to a cosmic reference frame. The first measurement of the dipole, consistent with the modern value and direction (3.358 ± 0.017 in direction $\alpha = 11.19^h$ at $\delta = -6.9^\circ$), was made by Ned Conklin from White Mountain in California (Conklin 1969), although there were lingering doubts about the accounting of foreground emission (Webster 1974). Clearer detections were later reported by Corey and Wilkinson (1976) and Smoot et al. (1977).

This velocity dipole, or "aether drift," was the first CMB anisotropy observed. However, it was clear that if cosmic structure grew from the Big Bang, then it had to leave an imprint as anisotropy in the CMB. The magnitude of the anisotropy was uncertain. The expansion rate, geometry, and composition of the universe were all poorly known. It was not clear just how large the foreground signals were relative to the anisotropy or if the reionization of the universe erased the anisotropy en route.

Before COBE, there were over 30 different experiments, including the RELIKT satellite mission, aimed at searching for the anisotropy, as detailed in Peebles et al. (2009). Many experiments yielded multiple observations in different configurations. The search took place over a range of angular scales and at many frequencies. As the bounds tightened, theories of structure formation evolved. There was a remarkable and steady advance that drove the development of new detector technologies: masers, heterodyne systems, bolometers, and new ways of observing the cosmos. Observations were made from balloons, airplanes, and on the ground from Saskatoon, Canada to the South Pole. At last, in 1992, almost 30 years after the discovery of the CMB, the anisotropy was discovered by the DMR instrument aboard the



■ Fig. 13-12

The COBE-DMR 4-year, 53 GHz sky map. The *top panel* shows the map with the dipole anisotropy included. This contribution is understood to arise from the proper motion of the Solar System barycenter with respect to the rest frame of the CMB, with a velocity of 369.3 km/s. The *bottom panel* shows the map with the best-fit dipole anisotropy subtracted off. The *red band* is microwave emission from the Milky Way, but the structure above and below the Galactic plane is CMB anisotropy that originates at the epoch of last scattering (The figure is adapted from Bennett et al. (1996))

COBE satellite (Smoot et al. 1992). The sky map from DMR's most sensitive 53 GHz band is shown in ➋ *Fig. 13-12*. By that time, the measured anisotropy level was somewhat higher than expected from theory. The interpretation paper, from Wright et al. (1992), reports "The observed anisotropy is consistent with all previously measured upper limits and with a number of dynamical models of structure formation. For example, the data agree with an unbiased cold dark matter (CDM) model with $H_0 = 50$ km/s Mpc and $\delta M/M = 1$ in a 16 Mpc radius sphere. Other models, such as CDM plus massive neutrinos [hot dark matter (HDM)], or CDM with a nonzero cosmological constant are also consistent with the COBE detection and can provide the extra power seen on 5–10,000 km/s scales."

By the time of the COBE-DMR discovery, receiver technology had advanced to the point where the anisotropy could, in principle, be detected in one night of observation. Using data taken from before the COBE launch, the FIRS team soon confirmed the COBE discovery as shown in ➋ *Fig. 13-13*.

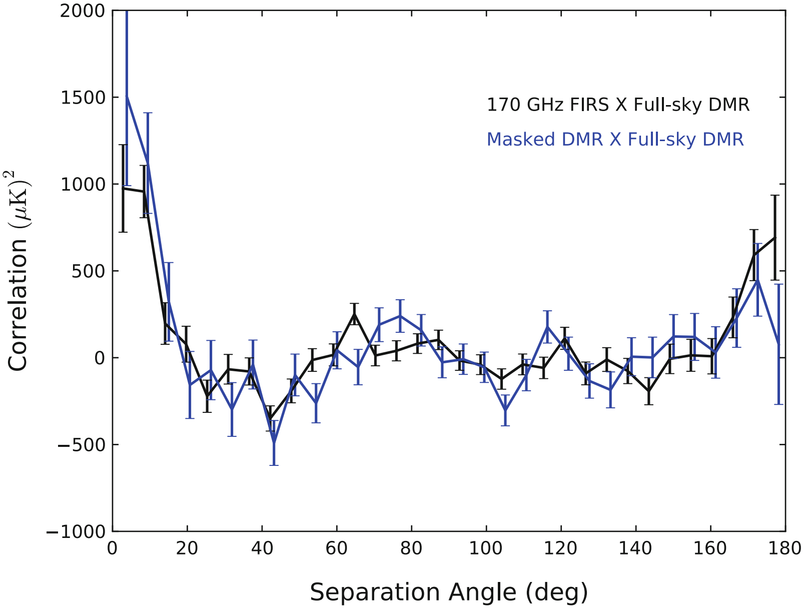


Fig. 13-13

The cross-correlation between the 170 GHz FIRS survey map and the COBE-DMR "Fit Technique" reduced galaxy map (*black*), compared to the autocorrelation of the DMR map (*blue*). FIRS covered roughly a quarter of the sky in a one-night balloon flight. Both estimates use the same portion of the sky to facilitate the comparison and excise data with $|b| < 15°$. The similarity is striking: the fluctuations in DMR were also present in FIRS. The uncertainties are correlated between angular bins, so assessing statistical significance is subtle, but the confirmation by FIRS is highly significant (The figure is adapted from Ganga et al. (1993))

4.2 From COBE to WMAP

The COBE-DMR experiment had relatively coarse angular resolution: 7° FWHM. The casual horizon size at decoupling, when the CMB photons were last scattered, is only 1.2°[1], so DMR was insensitive to causal physical processes occurring on these subhorizon scales. In particular, there were predictions that acoustic waves in the primordial plasma could have a coherence that would impart distinctive structure in the CMB anisotropy on subdegree (subhorizon) scales. Thus the COBE detection opened eyes to the potential for what the anisotropy on smaller angular scales could tell us. There were many questions: How large was the foreground emission on these scales? Did reionization erase the anisotropy at intermediate angular scales? Were acoustic features present, or did the power spectrum have less ringing? Did cosmic strings play a role in structure formation? Were the fluctuations produced by an isocurvature process, rather than an adiabatic one?

[1] We follow the conventions in Komatsu et al. (2011) because they are clearly defined but note that the often used scaling of $\Delta\theta \sim 180°/\ell$ gives different results.

If the primordial fluctuations were adiabatic, meaning that the constituent species of matter and radiation were perturbed in phase by the source mechanism, then the small-scale CMB anisotropy should exhibit a series of "acoustic peaks." Measuring the spacing and amplitude of the peaks would allow one to infer a number of fundamental cosmological parameters: the geometry and age of the universe, the abundance of baryonic and dark matter, and the slope of the primordial power spectrum of dark matter fluctuations. The fact that COBE-DMR was able to measure CMB anisotropy above the level of the galactic foregrounds on large angular scales gave the community hope that extracting the full potential of the CMB was within reach.

It is not possible to do justice to all the experimental work that took place in the decade between COBE and WMAP. There were roughly 25 independent efforts using a host of technologies and methods, all discussed in Peebles et al. (2009). The excitement of the science drove advances in detector and receiver technology. Over the decade, there were no major missteps, just a steady progression of more and more precise measurements. Immediately after COBE, there was a flurry of detections: the first measurement of the degree-scale power spectrum came from Saskatoon (Netterfield et al. 1997), while the existence and position of the first acoustic peak was identified by Miller et al. (1999), Knox and Page (2000), and Mauskopf et al. (2000). A tremendous advance by the Boomerang (de Bernardis et al. 2000) and Maxima (Hanany et al. 2000) teams revealed the CMB landscape with high precision. The next few acoustic peaks then revealed themselves over the next few years (Netterfield et al. 2002; Halverson et al. 2002; Ruhl et al. 2003) as did the Silk damping tail (Readhead et al. 2004; Kuo et al. 2004). The status of CMB anisotropy measurements just prior to WMAP is shown in *Fig. 13-14*.

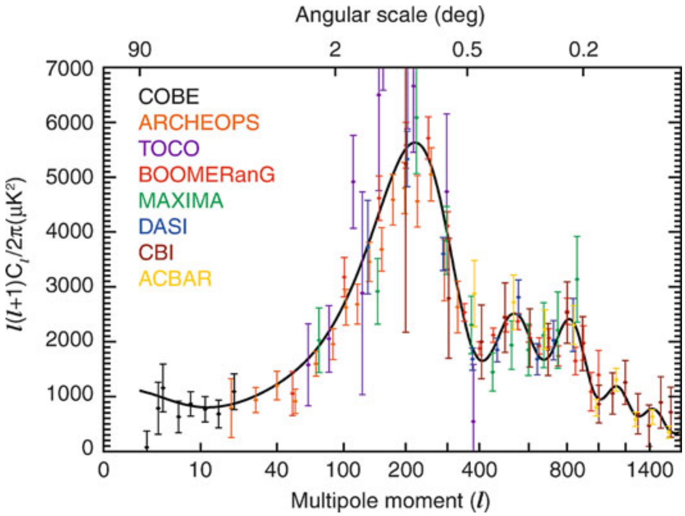


Fig. 13-14

Compilation of CMB power spectrum measurements immediately preceding the first WMAP results. The *black curve* shows the best-fit ΛCDM model from the first-year WMAP data for comparison. On average, the pre-WMAP data agree well with the WMAP power spectrum. The references for the previous data are as follows: COBE (Tegmark 1996); ARCHEOPS (Benoît et al. 2003); TOCO (Miller et al. 2002); BOOMERANG (Ruhl et al. 2003); MAXIMA (Lee et al. 2001); DASI (Halverson et al. 2002); CBI (Pearson et al. 2003); ACBAR (Kuo et al. 2004) (Figure from Hinshaw et al. (2003))

5 WMAP

In 1995, following the COBE-DMR discovery of CMB anisotropy, and contemporaneous with the experimental efforts noted above, the Microwave Anisotropy Probe (MAP) mission was proposed to measure CMB fluctuations with greater precision and accuracy than was possible with the DMR. Specifically, MAP was tailored to have an angular resolution of better than 1/3° in order to resolve the CMB's acoustic oscillation features expected to be present at subhorizon distance scales (about 2° across the sky at the distance of the last scattering surface). If present, the features would allow one to deduce detailed information about the shape, content, and age of the universe, among other things.

As MAP was being developed, the guiding principle was control of systematic errors (Bennett et al. 2003b). Among the design features incorporated into the mission were:

- A symmetric differential design
- Rapid large-sky-area scans
- Four switching/modulation periods
- A highly interconnected and redundant set of differential observations
- An L_2 orbit to minimize contamination from Sun, Earth, and Moon emission and allow for thermal stability
- Multiple independent channels
- Five frequency bands to enable a separation of galactic and cosmic signals
- Passive thermal control with a constant Sun angle for thermal and power stability
- Control of beam sidelobe levels to keep the Sun, Earth, and Moon levels <1 μK
- A main beam pattern measured accurately in-flight (using Jupiter)
- Calibration determined in-flight to the subpercent level (from the CMB dipole and its modulation from MAP's motion)
- Low cross-polarization levels (below −20 dB)
- Precision temperature sensing at selected instrument locations

Because of these multiple cross-checks, the MAP data could be understood in great detail. As discussed, the precision, accuracy, and reliability made possible by this has set the foundation for the standard model of cosmology.

➋ *Figure 13-15* shows a side view of the MAP's differential observatory, indicating two lines of sight to the sky which are sensed by the differential receivers located in the focal plane assembly (FPA), directly underneath the telescope. The FPA itself, shown in the lower right, prior to its integration on the spacecraft, houses 10 "differencing assemblies" spanning frequencies from 23 to 94 GHz.

MAP was launched at 15:47 EDT on 30 June 2001 from Cape Canaveral, FL aboard a Delta 7425–10 rocket. The satellite executed a sequence of phasing loops to position itself for a lunar swingby en route to the second Sun-Earth Lagrange point, L_2. MAP commenced cosmological observations from L_2 on 10 August 2001. The first results from MAP – based on one year of data – were released in February 2003. A few months prior to the first release, one of the leading figures in CMB research, and a founding member of the MAP team, Dave Wilkinson of Princeton University, passed away. The satellite was renamed the Wilkinson Microwave Anisotropy Probe (WMAP) in his honor. Subsequent data releases, based on 3, 5, and 7 years of data at L_2 were released in 2006, 2008, and 2010, respectively; and each release subsumes the earlier data and incorporates improved data processing algorithms. In this chapter we only discuss the most recent results based on 7 years of data.

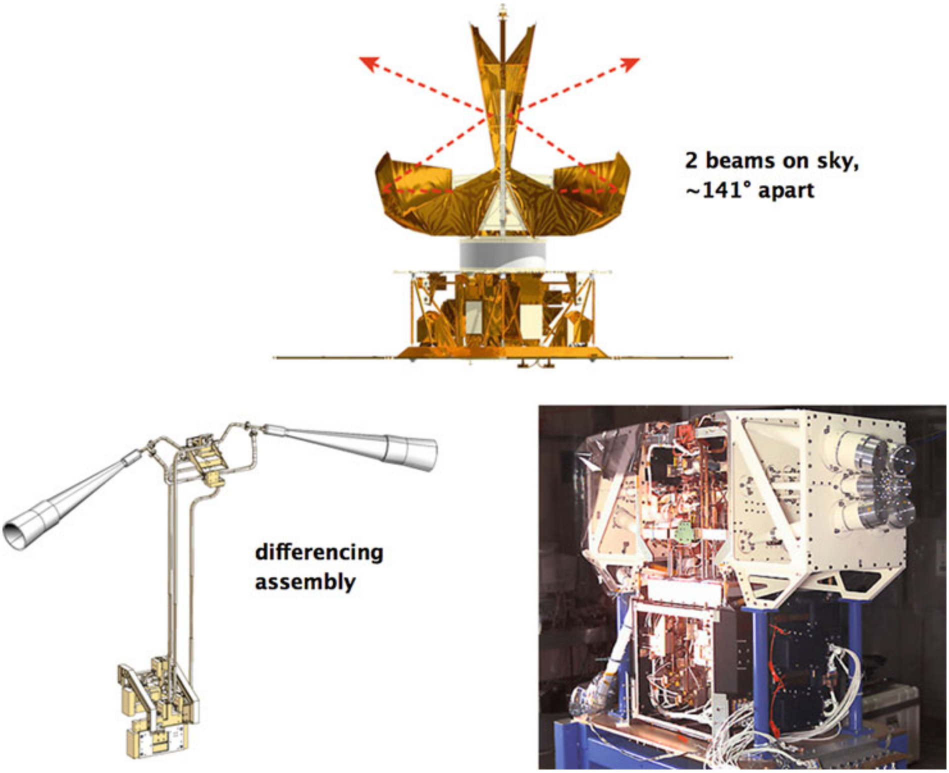


Fig. 13-15
The WMAP instrument consists of back-to-back Gregorian optics (*top*) that feed sky signals from two directions into ten 4-channel polarization-sensitive receivers called differencing assemblies (*bottom-right*). The HEMT amplifier-based receivers cover five frequency bands from 23 to 94 GHz. Each pair of channels is a rapidly switched differential radiometer (*bottom-left*) designed to cancel common-mode systematic errors. The signals are square-law detected, voltage-to-frequency digitized, and then downlinked

5.1 WMAP Sky Maps

The five frequency band maps produced from 7 years of WMAP observations at L2 are shown in *Fig. 13-16* (temperature anisotropy) and *Fig. 13-17* (polarization anisotropy). These full-sky maps are the primary product of the WMAP mission since they represent the most compressed form of the mission data one can produce without essential loss of information. The temperature maps show significant emission in the galactic plane due to synchrotron, free-free, and dust emission, but the CMB emission – seen clearly in the maps at high galactic latitude – is constant in thermodynamic temperature units. The polarization maps also show clear Galactic emission, primarily from synchrotron radiation, but the polarized CMB emission is only detected statistically – no specific feature seen in the polarization maps can be ascribed to the CMB.

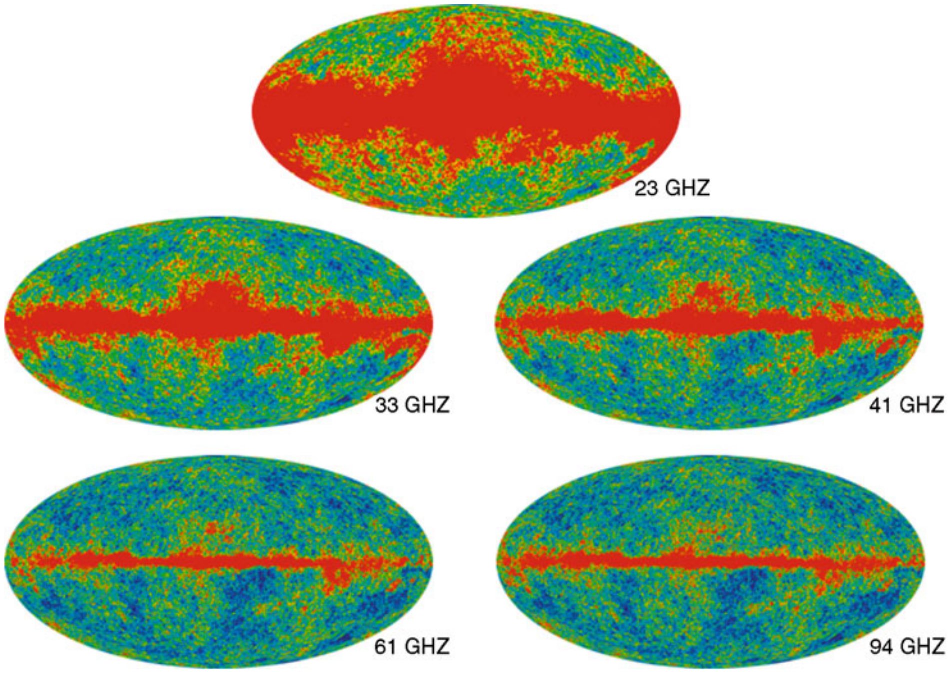


Fig. 13-16
The WMAP 7-year frequency band maps in galactic coordinates from 23 to 94 GHz. Note the strong frequency dependence of the galactic microwave emission and the constancy of the CMB fluctuations at high galactic latitude

The frequency spectrum of the Galactic emission is very different from the CMB. *Figure 13-18* shows the frequency spectra of the known galactic emission components, in antenna temperature units, compared to CMB anisotropy in the same units. The vertical range of each emission band indicates the *rms* signal level as a function of frequency if 77% (lower) or 85% (upper) of the sky is retained. The five WMAP frequency bands, which are each ~20% wide, span the minimum of the galactic foreground window. There are a variety of techniques that can be employed to separate the CMB component from the Galactic emission, each with their own advantages and disadvantages. Conceptually, the simplest approach is to form a linear combination of the frequency bands in such a way as to cancel signals with a Galactic spectrum while preserving signal with a CMB spectrum. This approach has been taken by the WMAP team to produce the "Internal Linear Combination" (ILC) CMB map shown in *Fig. 13-19* (Bennett et al. 2003a; Gold et al. 2011).

The CMB map depicted in *Fig. 13-19* looks like random noise, and in a very specific sense, that is exactly what it is. There is no theory of cosmology that predicts any specific feature in the CMB, only the statistical properties of the fluctuations in the map. As a result, we must analyze the map using statistics that can be compared to theoretical models. Perhaps the most remarkable feature of the CMB map is that there is a preferred angular scale in the data. One way to see this is via a stacking analysis: take every isolated hot spot (local maximum) in the map and excise a postage stamp image of the data in its vicinity. Now coadd the postage stamp images

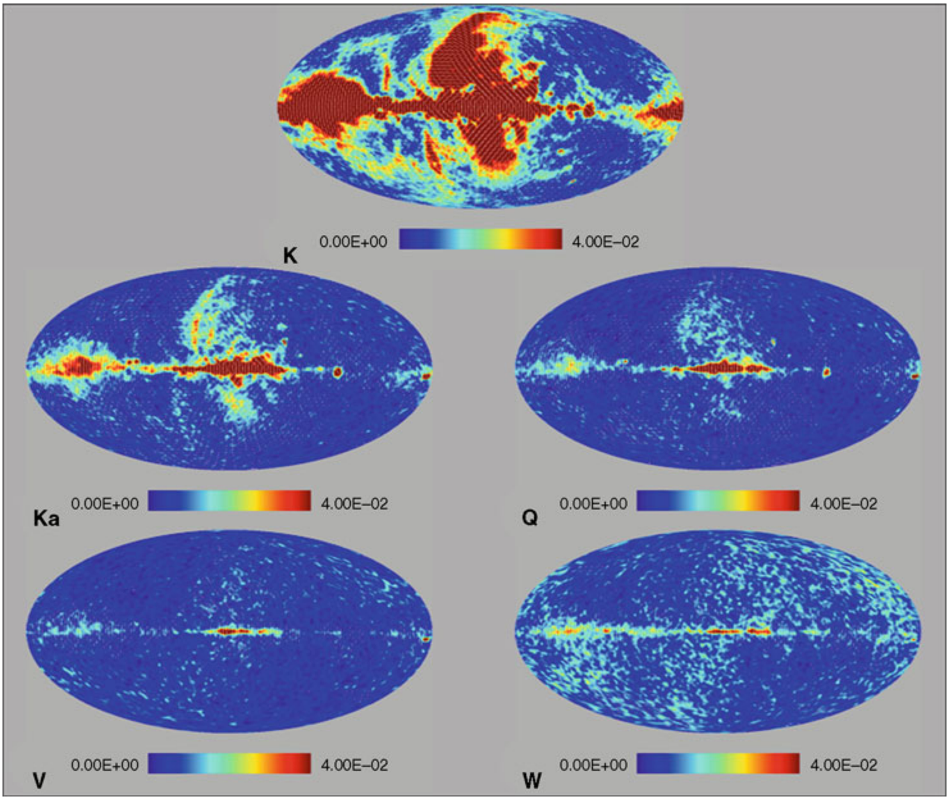


Fig. 13-17
The WMAP 7-year frequency band polarization maps from 23 to 94 GHz. The maps are dominated by galactic synchrotron and dust emission

with each one centered on the location of its respective temperature peak. Repeat the procedure for the locus of cold spots. The results of this stacking are shown in *Fig. 13-20*. The remarkable feature in the temperature data (left panels) are the concentric rings 1.2° from the central hot and cold spots. These features are the remnants of acoustic waves that propagated in the primordial plasma for 400,000 years until the electrons and protons "recombined" in the cooling plasma to form neutral gas which could no longer support acoustic waves. From this time on, CMB photons have propagated across the universe largely unimpeded, carrying with them the image of the last scattering or decoupling epoch, and all the information that it encodes.

5.2 WMAP Angular Power Spectrum

As noted above, cosmological models do not predict specific features in the CMB anisotropy, only their statistical properties. The simplest models predict that the CMB fluctuations are Gaussian distributed with random phase. We define the precise meaning of this shortly, but

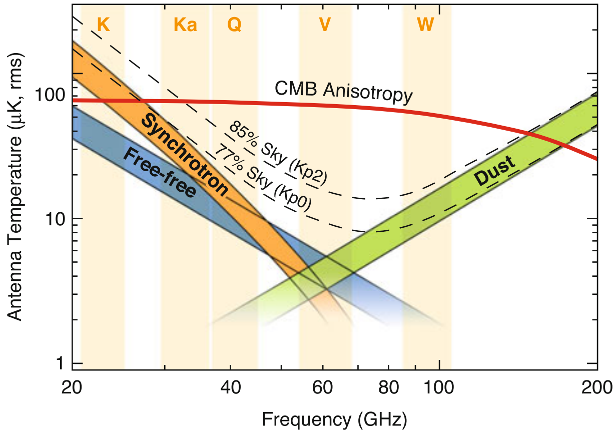


Fig. 13-18
The frequency spectra of CMB and galactic foreground emission in units of antenna temperature. WMAP observed in five bands because the foreground emission from our own Galaxy has a different frequency spectrum than the CMB (which is constant in the units used by WMAP), and the multifrequency data enables a separation of the two components

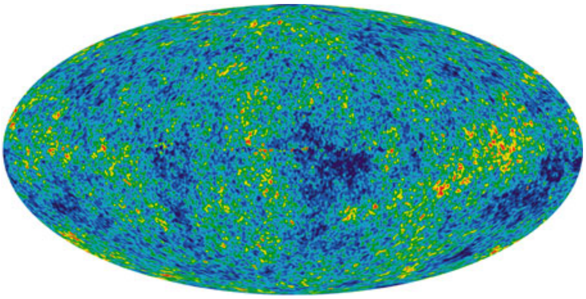


Fig. 13-19
The WMAP 7-year CMB map formed from an internal linear combination ("ILC") of the five frequency band maps, in such a way as to null emission with galactic frequency spectra

we note first that a Gaussian, random-phase field can be completely characterized by its 2-point correlation function, or equivalently, its angular power spectrum. Higher-order moments can be uniquely related to the 2-point (variance) statistics.

True for many analyses, it is convenient to expand the temperature anisotropy map, $T(\mathbf{n})$, in the basis of spherical harmonic functions, $Y_{\ell m}(\mathbf{n})$,

$$T(\mathbf{n}) = \sum_{\ell,m} a_{\ell m} Y_{\ell m}(\mathbf{n}). \tag{13.2}$$

(The inclusion of the monopole and dipole terms ($\ell = 0,1$) in this expansion are a matter of convention. Processes associated with these terms are distinct from those of the higher-order anisotropy. These contributions are discussed earlier in this article.) The leading explanation for the origin of the $\ell > 1$ anisotropy is that it arises from perturbations generated during inflation.

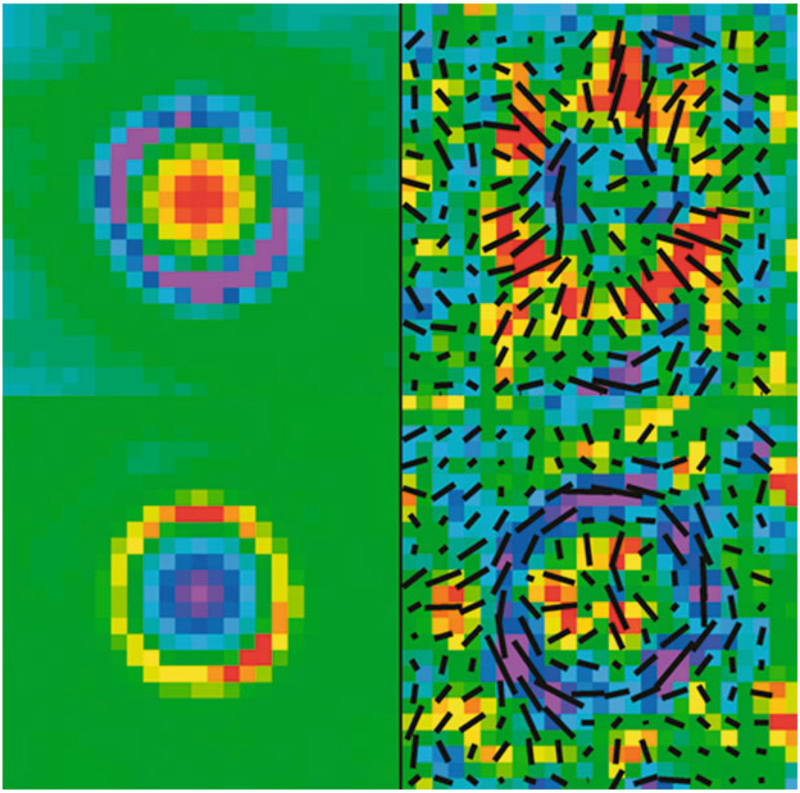


Fig. 13-20

The WMAP 7-year temperature and polarization maps stacked by location of hot spots (*top panels*) and cold spots (*bottom panels*). The temperature data are on the *left*, the polarization on the *right*. Each of the four panels represents a 5° × 5° square of sky. As discussed in the text, these images display the remnants of acoustic oscillations in the primordial plasma

We cannot predict the map $T(\mathbf{n})$ (or, equivalently, the $a_{\ell m}$ coefficients) but rather only the statistical properties of the $a_{\ell m}$. If the fluctuations are indeed Gaussian, distributed with random phases, then the $a_{\ell m}$ distribution is completely specified by its angular power spectrum,

$$C_\ell \equiv \langle |a_{\ell m}|^2 \rangle, \tag{13.3}$$

where the angle brackets indicate ensemble average, in this case over an ensemble of widely separated cosmic observers, each of whom samples a statistically independent realization of the cosmic fluctuation field. The primary goal of anisotropy measurements is to estimate the underlying power spectrum as accurately as possible from the data since this is the observable that allows one to constrain cosmological parameters.

In practice, since we have only one sample of the cosmic fluctuation field on our own surface of last scattering, our ability to infer the parent spectrum, C_l, is limited by "cosmic variance." Conceptually, if we have a full-sky anisotropy map, we can invert (13.2) to obtain the measured $a_{\ell m}$ and use them to estimate the power spectrum observed in our sky,

$$C_\ell^{\mathrm{sky}} = \frac{1}{2\ell+1} \sum_m |a_{\ell m}|^2. \tag{13.4}$$

Since each multipole, ℓ, has only $2\ell + 1$ independent m modes, the cosmic variance uncertainty inherent in the estimator C_ℓ^{sky} is

$$\Delta C_\ell = C_\ell \sqrt{\frac{2}{2\ell+1}}. \tag{13.5}$$

For the quadrupole ($\ell = 2$) power, this implies an irreducible uncertainty of 63%; at high ℓ, the fractional uncertainty falls as $\ell^{-1/2}$, becoming subdominant to instrument noise at suitably high ℓ. As of this writing, the measured spectrum uncertainty is limited by cosmic variance up to $\ell = 550$. The C_ℓ measurement by the Planck satellite is expected to be cosmic variance limited to $\ell \approx 1{,}500$.

Readers more comfortable with position space than Fourier space may prefer to characterize the fluctuation properties by the angular correlation function, $C(\theta)$. This is related to the angular power spectrum by a Legendre transform

$$C(\theta) = \frac{1}{4\pi} \sum_\ell (2\ell+1)\, C_\ell \, P_\ell(\cos\theta), \tag{13.6}$$

which can be estimated directly from a sky map by averaging the product of all temperature measurements separated by an angle θ,

$$C^{\mathrm{sky}}(\theta) = \langle T(\mathbf{n}) T(\mathbf{n}') \rangle, \tag{13.7}$$

where the average is taken over all direction pairs, $\mathbf{n}, \mathbf{n}'$ such that $\mathbf{n} \cdot \mathbf{n}' = \cos\theta$ (to within some finite bin size in θ).

In practice, real world effects such as instrument noise, systematic errors, and contamination from Galactic emission force us to adopt significantly more difficult methods for estimating the power spectrum and its uncertainty. However, due to clever work by a great number of researchers in the field, these hurdles have been largely overcome, and they do not qualitatively alter the simple description given above.

The angular power spectrum inferred from the 7-year WMAP data is shown in ➲ *Fig. 13-21*. The red curve is the best-fit ΛCDM model fit to the data (see ➲ Sect. 6). The error bars plotted with the data show the uncertainty due to WMAP's instrument noise, while the gray band centered on the ΛCDM curve indicates the cosmic variance associated with that model. We expect the power spectrum observed in our sky to fluctuate from the underlying smooth curve by an amount characterized by the gray band, even in the absence of instrument noise, and our inference of the best-fit model must account for this cosmic variance.

5.3 Beyond WMAP

Aside from WMAP, the flurry of experimental activity that preceded it has continued to the present day. ESA's Planck satellite is remapping the CMB sky with exquisite sensitivity and angular resolution more than twice that of WMAP's. Planck was launched on May 14, 2009, from Kourou, French Guiana. The cryogenic HFI instrument observed for over 2.5 years before its cryogens depleted in January 2012, while the H_2 sorption pump cooled LFI instrument

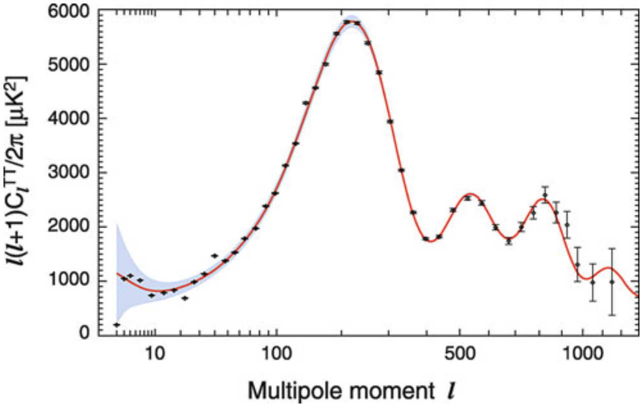


Fig. 13-21
The 7-year WMAP power spectrum. The *black points* with error bars show the data, the *red line* shows the best fit model, and the *gray swath* shows the cosmic variance, as discussed in the text. Although at first glance it may appear that some points are discrepant with the model, this is not supported by a statistical analysis. In particular, recall that roughly 1/3 of the data points should lie outside the cosmic variance swath

continues to observe as of this writing. The first cosmological results from Planck are expected in early 2013, and the community eagerly awaits them.

In the meantime, two flagship ground-based telescopes have each been deployed for a few years now: the South Pole Telescope is a 10-m observatory situated at the South Pole station, while the Atacama Cosmology Telescope (ACT) is a 6-m observatory deployed at an elevation of >5,000 m in the Atacama Desert in Chile. Both of these systems are equipped with state-of-the-art bolometric receiver systems that render them capable of measuring temperature anisotropy up to multipole moments of $l \sim 10{,}000$. Both SPT and ACT extend the reach of CMB observations well beyond that obtainable by WMAP (and even Planck, though the latter will have superior power spectrum sensitivity up to $l \sim 2{,}000$ due to its full sky coverage). A compilation of the WMAP, SPT, and ACT power spectrum measurements is shown in ➲ *Fig. 13-22*. The cosmological implications of these measurements are discussed next.

6 The Standard Cosmological Model

In this section, we interpret the CMB data and other observations in terms of a handful of cosmological parameters. We take as a starting point that the dynamics of the universe are governed by General Relativity and that the Cosmological Principle applies, namely that on suitably large scales, the universe is homogeneous and isotropic. In the context of General Relativity (a metric theory of gravity), the Cosmological Principle requires that the metric of space-time be of the Robertson-Walker form. This metric can have positive, negative, or zero spatial curvature, and the relative size of the universe can be described by a dimensionless scale factor, $a(t)$, whose dynamics are governed by the Friedmann equations: the Einstein equations applied to the Friedmann-Robertson-Walker metric. The specific behavior of $a(t)$ depends on the initial

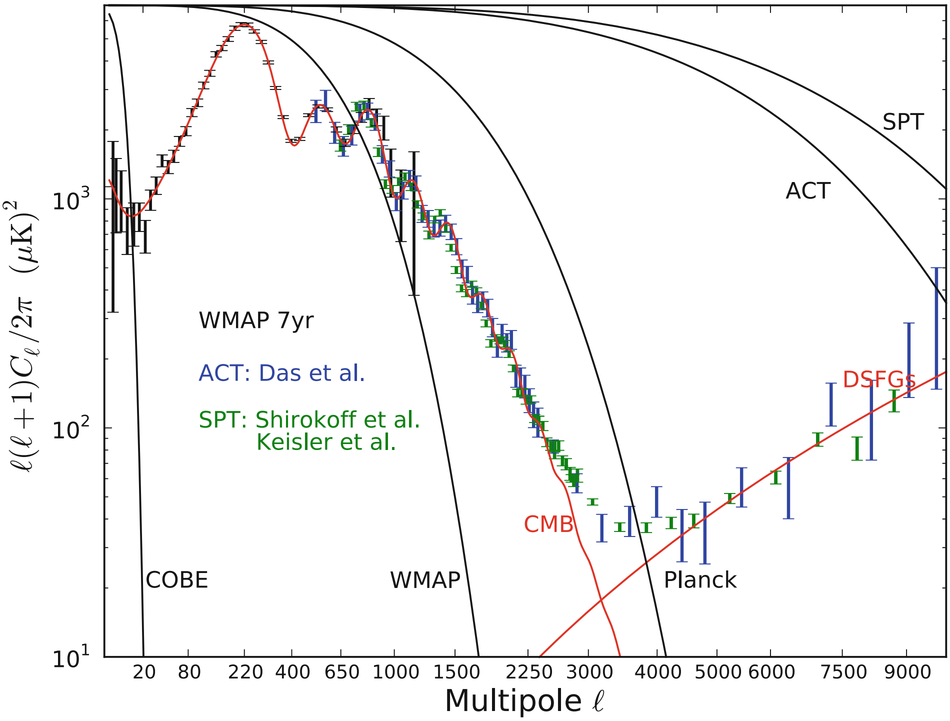


Fig. 13-22

A compilation of the most sensitive CMB power spectrum measurements published c. 2012 (Larson et al. 2011; Shirokoff et al. 2011; Das et al. 2011a; Keisler et al. 2011). The best-fit ΛCDM cosmological model (*solid red*) dominates the signal up to *l* ~ 3,000. The model power spectrum for ℓ > 3,000 is due to Poisson noise from confusion-limited dusty star-forming galaxies (DSFGs) observed at 150 GHz. The *black lines* indicate the angular resolution of various instruments, as measured by their window functions (the degree to which beam smoothing suppresses variance as a function of angular scale). The *l*-axis is scaled as $l^{0.45}$ to emphasize the middle range of the anisotropy spectrum. The increasing size of the WMAP uncertainties near *l* = 2 and 1,000 are due to cosmic variance and finite beam resolution, respectively. The cosmological implications of these data are discussed in Sect. 6

conditions and on the matter and radiation content of the universe, but to our knowledge, it has been monotonically increasing throughout time.

As discussed earlier, the blackbody spectrum of the CMB provides compelling evidence for a hot early universe that was in thermal equilibrium and which has expanded and cooled adiabatically. As we discuss later, the detailed expansion history is a topic of intense current research. The isotropy of the CMB brightness suggests an isotropic universe, consistent with the Cosmological Principle. While isotropy does not require homogeneity, isotropy without homogeneity would require that we occupy a special place in the universe, which seems untenable. Further, the observation that galaxies recede from us, on average, at a speed proportional to their distance, implies a homogeneous expansion; indeed, homogeneous expansion requires this proportionality. And, while some argue that there is no definitive evidence for a cutoff in the scale at which galaxies cluster, most workers in the field accept that at scales beyond ~1,000 Mpc, the clustering of matter becomes insignificant. Thus, throughout most of this chapter, we adopt the homogeneous and isotropic Hot Big Bang Model as our paradigm.

6.1 The Origin of Structure

The notion that the universe is homogeneous is a useful approximation, but it obviously breaks down at some point because there is clearly structure in our universe. On scales smaller than ~100 Mpc, the contrast between overdense and underdense regions exceeds unity; and on scales comparable to our Solar System, it far exceeds unity! In contrast, anisotropy in the CMB limits structure in the radiation field to be of order one part in 10^5, which implies corresponding limits on baryonic matter fluctuations on large scales, at early times. The basic paradigm for structure formation is that some process seeded fluctuations in the very early universe – cosmic inflation is the leading mechanism – and that these fluctuations grew in time by gravitational accretion to form the structure we see today. In fact, much of what we know about our universe comes from observing the evolution of cosmic structure over a range of scales. CMB anisotropy provides the cleanest probe of structure because it is weak, and therefore a product of simple linear physics, but later observations of structure as traced by galaxies are also crucial to constrain cosmological models.

It is beyond the scope of this chapter to cover perturbation theory in General Relativity, or the technology of N-body simulations that are required to study nonlinear structure. Rather, we will focus on what the current observations of structure tell us about cosmological models.

For the purpose of interpreting CMB anisotropy, the most important concept to understand is that of "baryon acoustic oscillations" (BAO). We imagine that some process like cosmic inflation imparted a spectrum of density perturbations to the universe at very early times – effectively $t = 0$ – and we wish to track how those perturbations evolve between then and the time when the CMB was decoupled 400,000 years later. The basic physics of acoustic oscillations was worked out in the late 1960s (Peebles and Yu 1970; Sunyaev and Zel'dovich 1970a; Silk 1968) and has been refined ever since, to the point where numerical codes that solve the Boltzmann transport equations can now predict matter and radiation perturbation spectra to an accuracy of better than 1% (Howlett et al. 2012).

If the initial fluctuations were "adiabatic," meaning that the various species of matter and radiation perturbed in phase, then an initial overdensity of baryonic matter will propagate outward as a sound wave front with a speed that is determined by the bulk modulus of the relativistic plasma, $c_s \approx c/\sqrt{3}$, where c is the speed of light. When the cosmic plasma becomes neutral at the epoch of recombination, the sound waves in the photon-baryon fluid "freeze out" due to the loss of photon pressure. The distance traveled by these wave fronts is called the sound horizon (or the BAO scale) and is known very precisely from the propagation velocity and the travel time, both of which only depend on the matter to radiation ratio, which is now quite well measured.

The effect of BAO on the radiation is to impart a feature in its 2-point correlation function or angular power spectrum. Specifically, we expect a coherent series of peaks in the power spectrum, corresponding to the harmonics of the propagating sound wave fronts. The detailed spacing and amplitude of these peaks contain a wealth of information about the geometry and contents of the universe. Indeed, the mere *existence* of these peaks is a remarkable triumph for theoretical cosmology. The fact that such structure was predicted long before it was considered observable is an excellent reminder that cosmology is a predictive science.

It is worth emphasizing this point once again. The relatively narrow width and coherence of the eight acoustic peaks seen in *Figs. 13-21* and *13-22* are an indication that the initial fluctuations were predominantly adiabatic. Specifically, the WMAP7 data limits out-of-phase “isocurvature” perturbations to be less than ~10% of the total signal. The fact that we are able to resolve these peaks at all is the key feature that makes the CMB such a powerful probe of cosmological parameters. The universe could have easily been much more complicated than it has turned out to be so far.

6.2 Geometry of the Universe

In the context of the General Relativity and the Friedmann-Robertson-Walker metric, a flat universe is an unstable fixed point. That is, any positive or negative spatial curvature present in the universe grows with time. In order for the universe to be within an order of magnitude of flat today – as measured by the total energy density, Ω_0 – it must have been extremely close to unity in the past. This fine-tuning problem, known as the flatness problem, was one of the motivating factors for cosmic inflation. Inflation addresses the flatness problem by invoking a period of accelerated expansion in the very early universe, which dilutes any existing spatial curvature to negligible levels. Generically, inflation predicts that the geometry should be should be flat to a part in 10^4.

CMB anisotropy probes geometry by measuring the angular size of the sound horizon, as probed by the position of the first acoustic peak in the power spectrum. If we know the distance to the last scattering surface, we can complete the triangle and determine the shape of light-like geodesics that span the space between us and the last scattering surface. Unfortunately, the distance to the last scattering surface is not known a priori, since it depends on the low-redshift normalization of the distance-redshift relation, that is, the Hubble constant. We illustrate this degeneracy in *Fig. 13-23*.

On the left are two hypothetical triangles with the same opening angle and sound horizon but different lengths. To first order the CMB is unable to distinguish between these two geometries. This is shown quantitatively in the right panel which shows allowable models that are consistent with the WMAP data. Each point in the panel represents a cosmological model that falls within the 95% confidence region of the WMAP7 data. The points are color-coded by the value of the Hubble constant required to produce the correct angular size of the sound horizon (as measured by the first acoustic peak position). The solid line with $\Omega_0 = \Omega_\Lambda + \Omega_m$ corresponds to a flat universe; points to the right of that line correspond to a closed universe. The WMAP7 data alone give $\Omega_0 = 1.0080^{+0.093}_{-0.071}$, but this jointly requires a low value for the Hubble constant, $H_0 = 53^{+13}_{-15}$ m s^{-1} Mpc^{-1}. If, in addition, we invoke the independent measurement of the Hubble constant by, for example, Riess et al. (2011) of $H_0 = 73.8 \pm 2.4$, the limits on curvature tighten to $\Omega_0 = 1.0023^{+0.056}_{-0.054}$, as shown by the blue contours on the right panel of *Fig. 13-23*. Remarkably, the Hubble constant measurement breaks the geometric degeneracy in the CMB right where the spatial geometry is flat (Euclidean). *Figure 13-24* illustrates the geometric degeneracy in the CMB in the context of the observable power spectrum. For the remainder of this chapter, we adopt a flat universe unless otherwise noted.

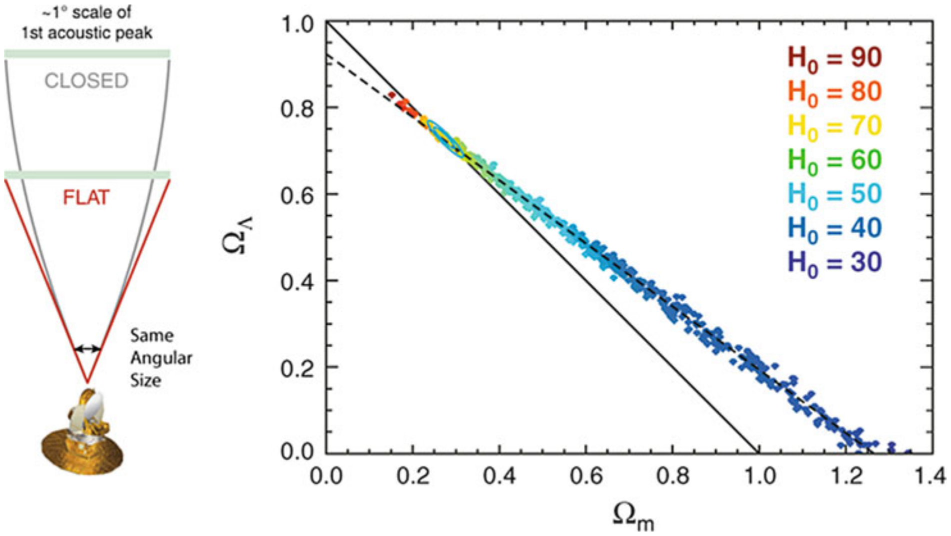


Fig. 13-23
The geometric degeneracy inherent in CMB anisotropy is shown as the *dashed line* – see text for details. The small blue contours in the *right panel* indicate that the combination of WMAP7 data and a recent measurement of the Hubble constant strongly favor a flat universe

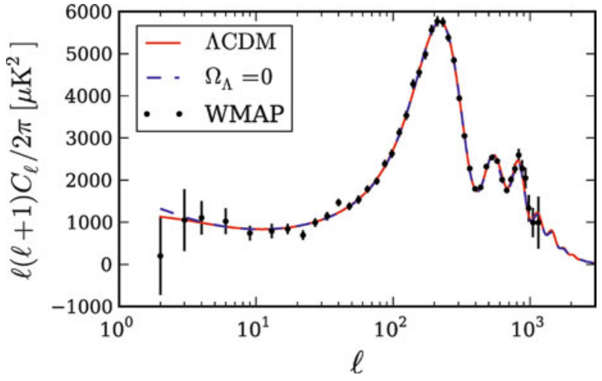


Fig. 13-24
The geometric degeneracy in the CMB angular power spectrum. The *red curve* is the standard ΛCDM model with $\Omega_\Lambda = 0.73$ and $\Omega_m \approx 0.26$. The *blue curve* corresponds to $\Omega_\Lambda = 0$ and $\Omega_m \approx 1.3$. The models are not significantly distinguishable with WMAP data *alone* (The figure is from Sherwin et al. (2011))

6.3 The Matter Content of the Universe

Given the interpretation of a flat universe, we turn to the matter and energy content required to produce $\Omega_0 = 1.0023^{+0.056}_{-0.054}$. The acoustic peak spectrum is sensitive to both the total matter density, Ω_m, and separately to the baryonic matter density, Ω_b. The sensitivity to both arises because the total matter density acts as a driving term to the acoustic oscillations, while the baryons – which participate in the oscillations – act as a small drag term. The odd and even harmonics in the acoustic spectrum correspond to modes that froze out at maximum compression and maximum rarefaction, respectively. Because of baryon drag, there is a small asymmetry between the compression and rarefaction in the BAO, which gives rise to a feature in the CMB power spectrum, namely that the even peaks will be suppressed relative to the odd ones as the fractional contribution of baryons to the total matter density, Ω_b/Ω_m, is increased.

If one restricts the cosmological model parameter space to the six parameters of the flat ΛCDM model: the physical baryon density, $\Omega_b h^2$; the physical cold dark matter density, $\Omega_c h^2$; the cosmological constant in units of the critical density, Ω_Λ; the slope of the primordial power spectrum, n_s; the optical depth of the reionized medium, τ; and the overall amplitude of the fluctuations, A, the 7-year WMAP data alone gives the following constraints on the matter densities: $100\Omega_b h^2 = 2.258 \pm 0.057$ and $\Omega_c h^2 = 0.1109 \pm 0.0056$, corresponding to $\Omega_b = 0.0449 \pm 0.0028$ and $\Omega_c = 0.222 \pm 0.026$ (Larson et al. 2011). If one combines WMAP with SPT data, the constraints tighten by 15–25% to $100\Omega_b h^2 = 2.22 \pm 0.042$ and $\Omega_c h^2 = 0.112 \pm 0.0048$ (Keisler et al. 2011).

Reducing the uncertainty on Ω_m is of interest for future tests of physics beyond the standard ΛCDM model since the matter abundance is coupled to CMB probes of neutrino physics and to the details of how dark energy affects the expansion history of the universe. The current uncertainty in Ω_m is dominated by uncertainty in the amplitude of the third acoustic peak, which is still limited by noise in the WMAP data, and by calibration uncertainty in the SPT and ACT data. The Planck satellite should produce a cosmic variance limited measurement of the third acoustic peak which will substantially improve the uncertainty in our knowledge of this key parameter.

6.4 The Age of the Universe

Given the assumption of a flat universe and the above determination of the matter and energy content (which explicitly assumes that $\Omega_\Lambda = 1 - \Omega_m$, ignoring the negligible present-day radiation density), we can derive the age of the universe by solving the Friedmann equations for the expansion history. The age is interpreted to be the elapsed time since $a(t) = 0$ (i.e., the "Big Bang"). This procedure also predicts the present-day expansion rate, $\dot{a}$ from which we can derive the Hubble constant, $H_0 \equiv (\dot{a}/a)_0$, where the subscript 0 means evaluated at the present time.

For the 7-year WMAP data, the age of the universe is determined to better than 1% precision: $t_0 = 13.75 \pm 0.11$ Gyr, with a predicted Hubble constant of $H_0 = 70.4^{+1.3}_{-1.4}$ km s^{-1} Mpc^{-1}. The latter is in good agreement with the independently determined (Riess et al. 2011) value noted above, $H_0 = 73.8 \pm 2.4$ km s^{-1} Mpc^{-1}. If we relax the assumption of flatness, but invoke the independent Hubble constant measurement to constrain the Friedmann solution, we derive a consistent age with an uncertainty of just under 2%, $t_0 = 13.86 \pm 0.26$ Gyr (Komatsu et al. 2011).

Note that these results assume that dark energy has the form of a cosmological constant. If the dark energy equation of state has a more "exotic" form that makes dark energy more significant in the early universe (as a fraction of the total energy density), the expansion history, and hence the age of the universe, could be somewhat different than the ΛCDM model predicts. There is currently no evidence to support such a model, but it cannot presently be ruled out.

6.5 Initial Conditions: The Inflationary Parameters

The inflation mechanism (Guth 1981; Linde 1982; Albrecht and Steinhardt 1982; Sato 1981) was proposed to resolve a number of problems with the "classical" Big-Bang Model. The main problems were the following: (1) "The monopole problem." In Grand Unified theories, there should be a sea of magnetic monopoles in the universe. None have been observed. (2) "The flatness problem." A slight deviation from flatness in the early universe is amplified by the cosmic expansion. The universe is so close to being flat today (➲ Sect. 6.2) that there seems to have been some kind of fine tuning at work to make the early universe extremely flat. (3) "The horizon problem." Two regions on the sky separated by more than $\theta_h = 1.2^\circ$ at the time of decoupling are causally disconnected given the standard expansion history. Yet the CMB is isotropic to a part in 10^5, implying that our entire observable universe evolved from a region that was once in thermal equilibrium, and thus in causal contact. Inflation resolves these problems by postulating a period of accelerated (likely exponential) expansion in the very early universe, possibly associated with a symmetry-breaking phase transition when the temperature of the universe was at the scale of grand unification, $T \sim 10^{16}$ GeV, $t \sim 10^{-35}$ s.

Inflation solves the above three problems as follows: if inflation takes place after the production of magnetic monopoles, their abundance will be diluted to negligible levels. Similarly, curvature decreases during accelerated expansion; if the overall growth factor experience during inflation is sufficient, the curvature will be reduced to negligible levels. Finally, during a period of accelerated expansion, the horizon becomes much larger than was previously inferred from "classical" decelerating expansion. In order for inflation to quantitatively solve these three problems, the universe must have inflated by a factor of at least 50–60 e-folds in the scale factor. This implies that the current energy density of the universe should be within about one part in 10^4 of the critical density, $\rho_c = 3H^2/8\pi G$. It will likely be a while before this level of measurement precision is reached.

A remarkable prediction of inflation is that all structure in the universe today was ultimately a product of quantum mechanical fluctuations in the microscopic preinflationary universe. During inflation, these fluctuations became "classical" density fluctuations on astrophysical length scales, and they subsequently evolved into the web of cosmic structure we observe today. The simplest physical models of inflation are driven by a single scalar "inflaton" field, and these models predict that the density fluctuations should be Gaussian distributed with random phases. In particular, if we have a density fluctuation field $\delta(\mathbf{x})$, and we Fourier transform it,

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x})}{\bar{\rho}} - 1 = \int d^3k\, \delta(\mathbf{k})\, e^{i\mathbf{k}\cdot\mathbf{x}}, \tag{13.8}$$

then the Fourier modes are Gaussian distributed with variance,

$$\langle \delta(\mathbf{k})\delta(\mathbf{k}') \rangle = \frac{2\pi^2}{k^3}\delta^3(\mathbf{k} - \mathbf{k}')P(k), \tag{13.9}$$

where $P(k)$ is the primordial power spectrum of density fluctuations. Single-field inflation models predict that the primordial spectrum should be well approximated by a power law over the range of scales we can observe with cosmic structure, $P(k) \propto k^{n_s - 1}$, where n_s is the spectral index of the fluctuation spectrum. Note that $P(k)$ gives the variance of fluctuations as a function of wave vector, k, which is roughly equivalent to angular scale on the fixed surface of last scattering.

Inflation further predicts the slope or "tilt" of the primordial power spectrum, as measured by n_s (Mukhanov and Chibisov 1981; Hawking 1982; Guth and Pi 1982; Starobinsky 1982; Bardeen et al. 1983).[2] As inflation proceeds, the inflaton potential energy and $P(k)$ decrease with time. Since the largest length scales inflate past the horizon first and "freeze out," the primordial spectrum is generically expected to have a "red" tilt, meaning that large scales should have a slightly higher primordial variance than small scales, that is, $n_s < 1$.

CMB data now indicate that the $P(k)$ is indeed lower at smaller scales. The 7-year WMAP data gives $n_s = 0.963 \pm 0.014$. The smaller scale SPT and ACT data expand the lever arm of physical scales over which to fit the primordial slope and give even more precise constraints. For example, the combination of WMAP and SPT gives $n_s = 0.9663 \pm 0.0112$. This is an amazing observation. Before inflation, Harrison, Peebles, and Zel'dovich posited $n_s = 1$ on the grounds of naturalness. Today, all indications are that n_s is less than unity at the 3σ level, which points to the need for a detailed physical model for the origin of fluctuations. Inflation provides one such class of models, which can now be put to the test with data. The idea that we can think of testing physics at the energy scales implied by inflation is nothing short of incredible.

Limits on deviations from a power law, as measured by the running index parameter, $\alpha = dn_s/d\ln k$ (Kosowsky and Turner 1995), show that α is within 2σ of zero.

Many models of inflation predict that there should be a background of primordial gravitational waves: propagating perturbations in space-time. In contrast to density fluctuations, which are associated with scalar perturbations in the space-time metric, gravity waves are associated with tensor perturbations. Single-field inflation models robustly predict that the amplitude of the gravity wave background is proportional to the energy scale at which inflation occurred (Baumann et al. 2009):

$$P_t(k) = \frac{2}{3\pi^2} \frac{V}{M_{\rm pl}^4}\bigg|_{k=aH}, \tag{13.10}$$

where $P_t(k)$ is the power spectrum of tensor perturbations (analogous to the density perturbation spectrum noted above), V is the energy scale of inflation, and H is the Hubble parameter during inflation.[3] The detection of primordial B-modes would impact far more than just cosmology. It would be the first example of an empirical connection between gravity and quantum field theory.

Observationally, the amplitude of the gravity wave background is parameterized by the "tensor to scalar" ratio,

$$r \equiv \frac{P_t(k)}{P_s(k)}\bigg|_{k=0.002\,\mathrm{Mpc}^{-1}}, \tag{13.11}$$

which is simply the ratio of tensor to scalar power spectra, evaluated at a particular wavelength. Both types of perturbations contribute to temperature anisotropy in the CMB, via the Sachs-Wolfe effect, but their relative contributions as a function of physical or angular scale are different, so they can be roughly distinguished on the basis of features in the temperature power spectrum. The current limit on r from CMB temperature measurements is $r \lesssim 0.2$ (95% cl) (Komatsu et al. 2011; Dunkley et al. 2011; Keisler et al. 2011). ➋ *Figure 13-25* shows the joint constraints on n_s and r from the 7-year WMAP data. In the limit that $r \lesssim 0.1$, the contribution of tensors to the temperature fluctuations becomes so small that they negligibly affect

[2] This prediction is not unique; for example, the cyclic model (Khoury et al. 2003) makes a similar prediction.

[3] Cyclic models predict no gravitational waves, and thus, if the waves are detected, these models will be ruled out.

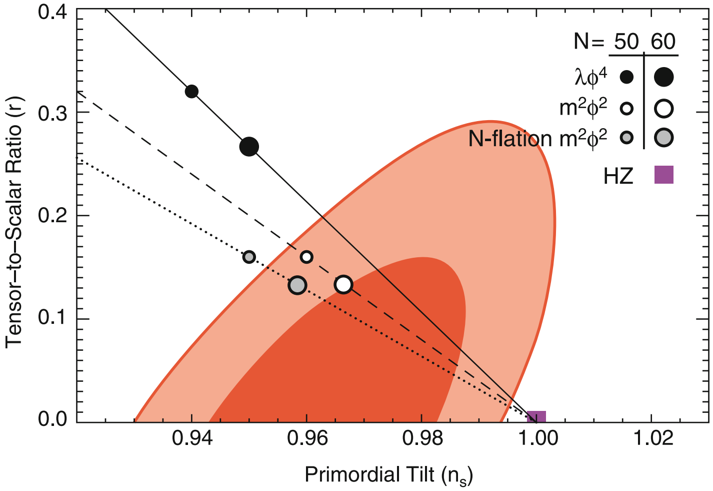


Fig. 13-25

The joint constraints on the scalar spectral index, *n_s* and the tensor (gravity wave) amplitude, expressed in terms of *r*. These constraints arise from the 7-year WMAP temperature data. (The figure is from Komatsu et al. (2011))

the power spectrum. Consequently, we need another channel for probing a weak gravitational wave background. Fortunately, there appears to be such a channel in the CMB polarization anisotropy.

6.6 Gravitational Waves and CMB Polarization

As noted above, both scalar and tensor perturbations contribute to temperature anisotropy in the CMB, via the Sachs-Wolfe effect. They both also contribute to polarization anisotropy in the CMB via Thomson scattering. However, the scalar symmetry of density perturbations restricts the polarization field these modes can produce. To understand this better, recall that polarization is a spin-2 field: any given pixel on the sky is characterized by a polarized amplitude and direction, the latter of which is invariant to rotations by 180°.[4] Two degrees of freedom are required to describe linear polarization, the most common of which are the two Stokes parameters Q and U (Hanany et al. 2012). In analogy to vector fields, which can be decomposed into gradient and curl components, an arbitrary polarization field can also be decomposed into so-called E-mode and B-mode components (Kamionkowski et al. 1997; Zaldarriaga and Seljak 1997). The symmetry of the scalar perturbations guarantees that they can only produce the gradient-like E-mode polarization. Thus, B-mode polarization provides a unique probe of propagating gravity waves at the epoch of recombination.

Since tensor perturbations are subdominant to scalar perturbations (see above) and since tensor perturbations can produce both E-mode and B-mode polarization, while scalars produce

[4] We specifically consider linear polarization here. Some astrophysical sources produce circular polarization, but we do not address those.

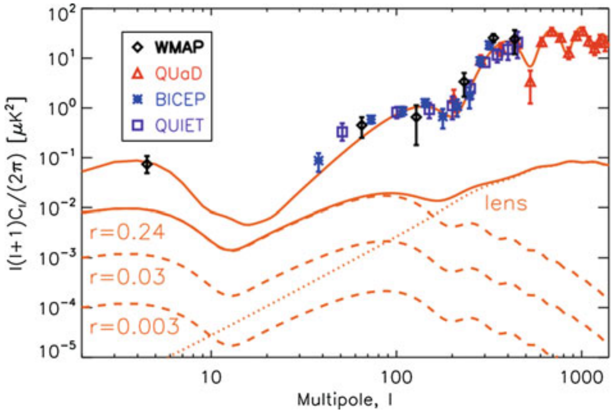


Fig. 13-26

Recent measurements of the CMB polarization at medium to large angular scales. The *top line* shows the E-mode spectrum in the same units as *Fig. 13-21*. The spectrum has already been quite well measured. The data are from WMAP (*diamonds*, Larson et al. (2011)), QUaD (*triangles*, Brown et al. (2009)), BICEP (asterisks, Chiang et al. (2010)), and QUIET (*squares*, QUIET Collaboration et al. (2011)). Primordial B-mode spectra are shown as *dashed lines* for different levels of primordial gravitational wave amplitude, *r*. The *top curve* r = 0.2 is disfavored at the 95% confidence level. The two peaks arise from reionization ($l \approx 5$) and decoupling ($l \approx 100$). The *dotted line* shows the B-mode spectrum induced by gravitational lensing of the primordial E-mode spectrum (*top*). The lensing signal is comparable to or larger than the r = 0.02 primordial spectrum for $l > 20$ and will be challenging to separate. Polarized foreground emission is not shown. Thus far, there is no detection of B-mode polarization in the CMB (The figure is from Katayama and Komatsu (2011))

only E-mode, it follows that B-mode polarization will be subdominant everywhere in the CMB. This makes them challenging to detect, but the rewards of doing so are high. At large angular scales, B-modes are produced by primordial gravitational waves. At all angular scales, they are produced by the gravitational lensing of E-mode polarization (Zaldarriaga and Seljak 1998). Lensing has the effect of displacing polarization “arrows,” which contaminates the symmetry of the E-mode signal induced by scalar perturbations at the surface of last scattering. We discuss these two phenomena in more detail below. *Figure 13-26* shows the E-mode and B-mode power spectra predicted and measured on medium to large angular scales.

As of this writing, there has been no observation of a significant B-mode signal in the CMB polarization anisotropy. The best direct limit on B-modes comes from the BICEP team, $r < 0.74$ (95% cl) (Chiang et al. 2010).

6.7 Building on the Standard Model

The standard ΛCDM model of cosmology, a flat, expanding universe dominated by dark matter and dark energy, with a nearly scale-invariant spectrum of density fluctuations, has been been around for roughly a decade. Observations that support the model have been made over a wide range of redshifts and wavelengths using many different cosmological observables, beyond just the CMB. To date, there are no observations in serious disagreement with the model. This is a remarkable state of affairs.

With the standard model as a foundation, there are many aspects of the model that remain to be understood. Did inflation occur in the very early universe? If so, what is the physical mechanism driving it? Is it related to spontaneous symmetry breaking? The most promising observational probes of inflation still reside in the CMB in the form of searches for B-mode polarization and non-Gaussianity in the temperature fluctuations. What is the nature of the dark energy that is driving the current acceleration of the universe? Present observations suggest a "vanilla" cosmological constant, but physicists are at a loss to understand its observed amplitude: they predict its natural value should be 10^{120} times higher than observed! Future observations of the expansion history of the universe may point to something other than a cosmological constant. What is the physics and astrophysics of cosmic structure formation? The former appear to be deeply tied to the physics of inflation; searches for primordial non-Gaussianity may be fruitful in this regard. Understanding the astrophysics is a key element of understanding how we came to be, and it may hold further clues about the nature of the dark matter.

On a different front, one may use the CMB to measure the amount of helium in the universe at the time of decoupling. This can be compared to the amount predicted by Big Bang nucleosynthesis at $z \sim 10^{10}$ and to the abundance observed today. Making sure these pieces of the puzzle fit together is an important step in cementing our understanding of Big Bang cosmology. And on yet another front, new measurements of the CMB hold the potential to constrain the sum of the neutrino masses. As discussed below, gravitational lensing of the CMB is providing a powerful new tool for probing the physics of our universe. There is much more to be done.

7 Anisotropy and Polarization Measurement Frontiers

The WMAP satellite is cosmic variance limited up to $\ell = 550$, and the Planck satellite is expected to be cosmic variance limited up to $\ell \approx 1{,}500$ as discussed above. In other words, assuming that the foregrounds have been properly accounted for and that any form of non-Gaussianity is negligible, all the information that can be extracted from the temperature anisotropy alone up to $\ell = 1{,}500$ will be present in the Planck maps. Planck will also measure the so-called E-mode polarization in the CMB and be cosmic variance limited on them to $\ell \approx 1{,}000$. The primordial B-modes are another independent observable that have yet to be detected. As we discuss below, there is at least one known cosmic source of non-Gaussianity. The source is the gravitational lensing of the CMB. Thus, there is an immediate motivation for pushing below the cosmic variance limit. In addition, there may be some form of primordial non-Gaussianity. Its detection would have a dramatic impact on our knowledge of how the universe began.

The two frontiers in the Planck era are the search for the large angular scale B-modes and the polarization and temperature anisotropies at fine angular scales, beyond Planck's resolution. It will be quite some time before the full sky is mapped again at Planck-like resolution with better sensitivity.

7.1 Large Angular Scale B-mode Experiments

Gravitational waves, as observed through the B-mode polarization, are a pristine probe of the early universe. This can be appreciated as follows: Quantum fluctuations in the primordial fields give rise to both scalar modes (variations in density as a function of position) and tensor modes (variations in strain – gravitational waves – as a function of position). Both scalar and tensor modes produce temperature and E-mode polarization anisotropies, but only tensor modes produce the B-mode polarization. Thus, the B-mode is distinctive. Although the gravitational waves are revealed to us through Thomson scattering off free electrons, they are unaffected by any cosmic process other than the expansion of the universe. They come to us directly from the inflation epoch.

Outside of experimental sensitivity, there are two astrophysics limitations to measuring B-modes. First, gravitational lensing of the E-mode signal masks the primordial B-mode signal for $\ell > 200$ as seen in *Fig. 13-26*. At $\ell \sim 100$, the lensed signal corresponds to $r \sim 0.02$. While it is possible to subtract the lensing, one pays a price in signal to noise. Secondly, the ultimate limit will be set by galactic foreground emission (Bock et al. 2006). Fortunately, models suggest that in the low-dust/low-synchrotron regions of sky, (Dunkley et al. 2009) we may reach $r \approx 0.02$ at 150 GHz before foregrounds become a limitation. At the largest angular scales, $\ell \lesssim 10$, polarized foreground emission dominates an $r \sim 0.02$ signal by over an order of magnitude (Page et al. 2007; Gold et al. 2011).

Measuring B-modes is challenging. Our confidence in any detection will be bolstered by multiple levels of redundancy, internal cross-checks, and agreement among experiments. The ability to identify and control systematic errors will ultimately determine the best approach (or approaches), which is not yet known. The set of experiments in *Table 13-2*, a snapshot of the current efforts focused on large angular scales, cover a wide range of technique.

Table 13-2

This information comes from a variety of sources (papers, conversations, presentations, web pages) and is intended only for a high-level comparison between efforts. The focal planes are different and not all have two detectors (D) per feed (F). ABS has an all-cryogenic cross-Dragone style telescope and observes from Chile. CLASS, also to be sited in Chile, will target circular as well as linear polarization. KECK/BICEP has observed for a number of years from the South Pole. QUBIC, the merging of the MBI (Timbie et al. 2006) and BRAIN (Charlassier for the BRAIN Collaboration 2008) efforts, is a novel bolometric interferometer that is anticipated to benefit from the same level of control of systematic errors enjoyed by its coherent predecessors. QUIET, a cross-Dragone, is based on coherent "polarimeters on a chip" as opposed to bolometric detectors. EBEX, PIPER, and SPIDER are balloon borne

Name	Res.	Freq. (GHz)	# Feeds/Dets
ABS (Essinger-Hileman et al. 2010)	0.5°	150	240F
CLASS (Marriage, 2012, private communication)	~1°	40, 90, 150	. . .
BICEP2/KECK (Sheehy et al. 2011)	0.7°	150	256F/5x256F
QUBIC (The QUBIC collaboration et al. 2011)	0.4°	150	400F
QUIET (QUIET Collaboration et al. 2011)	0.5°	40, 90	19F/90F
EBEX (Reichborn-Kjennerud et al. 2010)	0.2°	150, 250, 410	≈1, 500
PIPER (Eimer et al. 2010; Benford et al. 2010)	~0.5°	200, 270, 350, 600	5120D
SPIDER (Crill et al. 2008)	1.1°–0.4°	90, 145, 220	1856

7.2 Small-Scale Anisotropy, $\ell > 2{,}000$

Precise measurements of the small-scale anisotropy, $\ell > 2{,}000$, are a new frontier for CMB studies and a critical complement to the Planck satellite. Because of dramatic advances in detector and receiver technology, much of it driven by CMB-experimentalists, the sky is now being observed with arcminute-level resolution with cryogenic arrays of thousands of detectors. Thousands of square degrees of sky are being mapped with sensitivities measured in tens of microKelvin per square arcminute with special purpose telescopes: the Atacama Cosmology Telescope (ACT, Fowler et al. (2007) and Swetz et al. (2011)), located in Chile, and the South Pole Telescope (SPT, Ruhl et al. (2004) and Carlstrom et al. (2011)). Soon these receivers will be polarization sensitive. Additional instruments with a range of resolutions (e.g., POLARBEAR, The Polarbear Collaboration et al. (2010), POLAR-1, Kou, 2012, private communication) will soon come on line as well. In the not-too-distant future, one expects an order-of-magnitude improvement over current sensitivities.

The scientific questions that can be addressed with these measurements include: (1) What is the scalar spectral index and to what degree is its determination contaminated by foreground emission? (2) What is the sum of the neutrino masses and are there more than three relativistic species in the early universe? (3) Did the dark energy act differently before $z = 1$? (4) Where are the missing baryons? Big Bang nucleosynthesis and CMB-derived baryon densities are not in accord with the observational census. (5) Did the early universe have only Gaussian fluctuations? The discovery of primordial non-Gaussianity, perhaps from cosmic strings or multifield inflation, would revolutionize cosmology. (6) How did the universe evolve and how did cosmic structure form? Are there cluster of galaxies so massive and distant that they challenge the Lambda-dominated cold-dark-matter model of the universe?

The questions are addressed by investigating the CMB in a number of ways. For example, at small angular scales, one may think of the CMB as a backlight with precisely known statistical properties at a precisely known distance. This light is affected by structure between us and the decoupling surface. In the Sunyaev–Zel'dovich (SZ) effect, for example, the hot electrons in galactic clusters reveal their presence by scattering the CMB with a characteristic frequency signature. In another mechanism, mass concentrations throughout the universe gravitationally lens the CMB. This lensing can be determined by examining the correlations it imposes on the CMB. Through a rich set of lensing cross-correlations with other X-ray, optical, mm-wave, and radio surveys, we can probe the formation of structure. Through yet another mechanism, we can observe the decoupling process with the CMB polarization.

At angular scales corresponding to $\ell < 2{,}000$ the anisotropy may be thought of as a direct probe of the response of the CMB to perturbations laid down in the early universe as seen at the decoupling surface at $z = 1{,}090$. The fluctuations are a part in 10^5 of the background and thus well described with linear perturbation theory. The predictions for the CMB power spectrum may be computed with exquisite accuracy. The agreement between detailed predictions and precise measurements is the foundation for our faith in the standard model of cosmology as discussed above.

As one moves to finer angular scales, new phenomena are evident as shown in ➲ *Fig. 13-22*. In the structure formation process, aggregations of dark matter result in the formation of clusters of galaxies and galaxies. At 150 GHz, after accounting for discrete radio sources, the dominant term in the power spectrum is due to confusion noise from unresolved dusty star-forming galaxies (DSFGs). They formed at redshifts between $z \sim 1$–4 and carpet the sky. Their contribution can be minimized by observing at lower frequencies but at some point a similar background from unresolved radio sources becomes important. The DSFGs are generally described by a Poisson distribution, that is with C_ℓ is independent of ℓ, although there are correlations and clumping which lead to departures from this simple scaling (Bond 1996; Scott and White 1999; Viero et al. 2009; Hall et al. 2010; Addison et al. 2012).

Subdominant to the CMB and DSFGs at 150 GHz is the power spectrum from galactic clusters due to the Sunyaev–Zel'dovich (SZ) effects. There are two contributions. One is from the thermal effect in which the 10^7–10^8 K gas in clusters inverse Compton scatters CMB photons producing a unique frequency spectrum. When the cluster is viewed against the CMB, $T < T_{\rm cmb}$ for frequencies below 220 GHz, and $T > T_{\rm cmb}$ at higher temperatures. This effect was first observed in Birkinshaw et al. (1984). The second effect is from the Doppler shift of the CMB by the peculiar velocity of the cluster. In effect, the clusters act as moving mirrors. This effect was first observed by Hand et al. (2012) through stacking CMB measurements on the positions of many clusters.

Clusters are the largest gravitationally bound objects. The clusters that contribute to the power spectrum are being detected in blind surveys (Staniszewski et al. 2009; Hincks et al. 2010; Planck Collaboration et al. 2011b). They have masses greater than a few $10^{14} M_\odot$. However, roughly 50% of the amplitude comes from clusters with masses less than 2×10^{14} which will be difficult to detect in blind surveys. The SZ power spectra are produced from large simulations. The physical processes that determine the overall amplitude are all "sub grid" and must be modeled. Identifying the best description of clusters that explains X-ray, SZ, and optical data is an active area of research.

The SZ signal is approximately independent of redshift, and so clusters can be observed to great distances. The gas temperature goes as the depth of the gravitational well that led to the cluster formation and is nearly redshift independent. At high redshift, the cluster would naively appear dimmer, but back then the CMB was hotter and this compensates for the greater distance. Their mass function, $dN(>M)/dM$, that is, number distribution as a function of mass, is a sensitive probe of cosmology. Currently, the limiting factor for using clusters as cosmological probes is determining a precise mass. A typical error on the mass is 15%. Looking ahead, one anticipates the detection of over 10^5 clusters through their X-ray signature and roughly 10^4 clusters through their SZ signature. We can think of clusters as beacons positioned in a manner that lets us investigate the evolution of space-time. This will lead to a powerful check of the standard cosmological model and will undoubtedly tell us more about how cosmic structure forms and about the role of various cosmic constituents.

In the following sections we discuss three aspects of the small-scale anisotropy that will become increasingly important in the next few years. We discuss polarization and lensing, new modalities for CMB observations, and what we might learn from them. We conclude with a discussion of assessing the sum of neutrino masses through observations of the CMB.

7.2.1 Small Angular Scale Polarization

The CMB polarization is produced by different mechanisms than the temperature anisotropy and thus is sensitive to different physical processes. At small angular scale, the polarization probes the evolution of the decoupling. The polarization was predicted by Rees (1968) and Basko and Polnarev (1980) and first measured by Kovac et al. (2002). The signal is generated by Thomson scattering of a local quadrupolar radiation pattern by free electrons. The scattering of the same quadrupolar pattern in a direction perpendicular to the line of sight to the observer has the effect of isotropizing the quadrupolar radiation field. The net polarization results from a competition between these two effects. Basko and Polnarev (1980) show that the ratio of the polarization anisotropy ($E_{\rm rms}$) to the temperature ($T_{\rm rms}$) signal in a flat cosmology is given by

$$\frac{E_{\rm rms}}{T_{\rm rms}} = \frac{\int_0^\infty \left[e^{-0.3\tau(z')} - e^{-\tau(z')}\right]\sqrt{1+z'}dz'}{\int_0^\infty \left[6e^{-\tau(z')} + e^{-0.3\tau(z')}\right]\sqrt{1+z'}dz'}, \tag{13.12}$$

where $\tau(z) = c\sigma_T \int_0^z n_e(z')dz'(dt/dz')$ is the optical depth. Here, σ_T is the Thomson cross section, c is the speed of light, and n_e is the free-electron density. Using a typical optical depth (e.g., Peebles 1968b; Zel'dovich et al. 1969), one finds $E_{\rm rms} \approx 0.05\ T_{\rm rms}$. The difference in brackets in the numerator sets the range in z over which polarization is generated. For example, if the decoupling epoch entailed an instantaneous transition from an extremely high optical depth ($\tau >> 1$) to transparency ($\tau = 0$), there would be no polarization signal. Thus, the polarization is produced at a particular time. In contrast, the processes that lead to the temperature anisotropy take place over a much longer time.

At decoupling, the polarization producing quadrupole results from velocity gradients in the flow of the primordial plasma. More specifically, in the rest frame of an electron in such a flow, the radiation background has a quadrupolar pattern proportional to the velocity gradient, $\nabla\vec{v}$, and the mean free path between scatterings, λ. Just before decoupling, $z > z_{dec}$, the photons are tightly coupled to the electrons, and λ is small. Thus, the polarization is small. As decoupling proceeds, λ increases and the quadrupole magnitude increases. The process is cut off at lower redshift because the optical depth drops so rapidly. In the context of inflationary cosmology, Harari and Zaldarriaga (1993) show that in Fourier space the polarization signal is $\propto\ kv\Delta$, where k is the wavevector and $\Delta \approx \lambda$ is the width of the last scattering surface.

➲ *Figure 13-27* shows a snap shot of the $\ell > 200$ polarization spectrum along with predictions for what Planck will measure. The plot also shows the levels achievable with the current generation of polarization-sensitive experiments aimed at high ℓ. With E-mode polarization, one may probe the damping tail more deeply than with temperature and therefore better measure n_s. In addition, since the E-mode polarization arises from different physical processes, alternative models can be tested. For example, current power spectrum data indicate that temperature fluctuations are adiabatic to within 9% and 2% for axion and curvaton-type dark matter (Komatsu et al. 2009; Sollom et al. 2009). Small-scale temperature and polarization data will provide a new test of these alternative models.

As discussed above, our most direct probe of the infant universe is the scalar spectral index, n_s, and its change with scale. The formal accuracy on n_s from the Planck satellite is 0.5%. However, our confidence in the result will depend on detailed knowledge of the transition from the linear regime (primary CMB) to the nonlinear regime (secondary CMB). This transition can only be measured through the fine-scale anisotropy. In addition, we will want to be certain that n_s is not being affected by foreground emission, point sources, or low levels of secondary anisotropies. This is best done through the polarization.

The promise of the polarization as a new probe of cosmology is noteworthy. The CMB polarization was first observed just a decade ago. Now it looks like the cleanest measure of the CMB power spectrum over an appreciable range in ℓ may come not from the temperature but from the polarization.

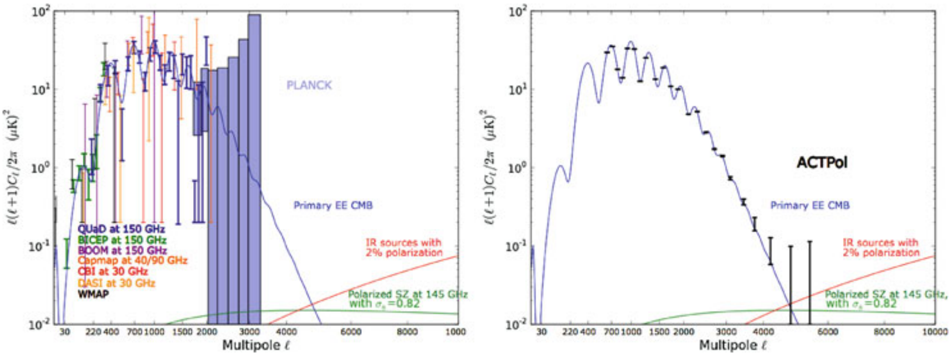


Fig. 13-27

***Left:* The current state of measurements of the E-mode polarization spectrum plotted over the best-fit ΛCDM model. Data are from QUaD (Brown et al. 2009), BICEP (Chiang et al. 2010), Boomerang (Montroy et al. 2006), CAPMAP (Bischoff et al. 2008), CBI (Sievers et al. 2007), DASI (Kovac et al. 2002), and WMAP (Larson et al. 2011). The Planck-projected errors are shown as *blue boxes*. The foregrounds are conservative estimates for IR point sources and a 1% net polarization of the SZ effect. *Right:* Projection of the errors from polarization-sensitive receivers on ACT and SPT. Note that the CMB is significantly more polarized than the foreground emission, permitting a detailed investigating of the damping tail and lensing**

7.2.2 Lensing of the CMB

One of the forefronts of CMB observations is the lensing of the CMB by matter fluctuations between us and the surface of last scattering. We see in many optical telescope images magnificent pictures of distant galaxies being gravitationally lensed by a large cluster of galaxies between us and the very distant galaxies. *Figure 13-28* shows an example. One thing to note is that the lensed objects are highly distorted. They can be magnified and elongated.

For the CMB, one replaces the distant galaxies by the CMB itself, and instead of the intervening galaxy, one has a universe full of matter fluctuations. The same power spectrum, $P(k)$, that gives rise to the temperature anisotropy also gives rise to the clumpiness of the universe between us and the surface of last scattering. In place of seeing beautiful images of distorted galaxies, the hot and cold spots in the CBM are subtly distorted but in such a manner as to not change surface brightness.

CMB lensing is described in a number of seminal papers (e.g., Blanchard and Schneider 1987; Seljak 1996) and in several recent review articles (e.g., Lewis and Challinor 2006; Smith et al. 2009). It is quantified as follows. When we measure the temperature at a given spot we find

$$T(\boldsymbol{n}) = T^u(\boldsymbol{n} + \vec{d}(\boldsymbol{n})), \tag{13.13}$$

where T^u is the unlensed temperature distribution, $\boldsymbol{n}$ is the direction of the observation, and $\vec{d}$ is the deflection field transverse to the direction. The deflection field may be written as the gradient of a potential, $\vec{d} = \nabla_{\boldsymbol{n}}\phi$. In turn, the lensing potential is related to the distribution of the gravitational potential throughout the universe, Ψ, as

$$\phi(\boldsymbol{n}) = -2\int_0^{z_{\rm dec}} \frac{\Psi(z, D(z)\boldsymbol{n})}{H(z)}\left(\frac{D(z_{\rm dec}) - D(z)}{D(z_{\rm dec})D(z)}\right)dz. \tag{13.14}$$

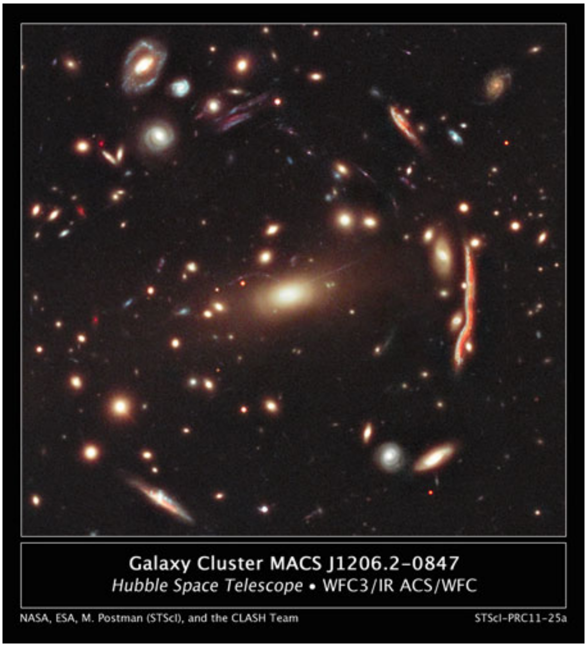

Galaxy Cluster MACS J1206.2-0847
Hubble Space Telescope • WFC3/IR ACS/WFC
NASA, ESA, M. Postman (STScI), and the CLASH Team
STScI-PRC11-25a

Fig. 13-28
A Hubble Space Telescope image of the galaxy cluster MACS J1206.2-0847 from the CLASH survey. The distortion of the distant galaxies is clearly evident. Such a cluster also distorts the CMB although the observations are not as dramatic

For CMB lensing, the source is fixed at the decoupling surface at a redshift of $z_{\rm dec}$. The quantity $D(z)$ is the comoving distance to an object at redshift z. Inside the large parentheses, one sees the familiar expression for the geometry of an object at $z_{\rm dec}$ being deflected by a lens at $D(z)$. The integral sums up contributions over Ψ, the source of the deflections. The power spectrum of ϕ is given by

$$C_\ell^{\phi\phi} = \frac{8\pi^2}{\ell^3}\int_0^{z_{\rm dec}} \frac{D(z)}{H(z)}\left(\frac{D(z_{\rm dec}) - D(z)}{D(z_{\rm dec})D(z)}\right)^2 P_\Psi(z,k)dz, \qquad (13.15)$$

where $P_\Psi(z,k)$ is the power spectrum of the gravitational potential as a function of redshift and k-vector, with $k = (\ell + 1/2)/D(z)$. In the literature, one sees a number of expressions for the lensing power spectrum. For the power spectrum of the deflection angle, $C_\ell^{dd} = \ell^2 C_\ell^{\phi\phi}$. For numerical work, the convergence, $\kappa = \frac{1}{2}\nabla \cdot \vec{d}$, is particularly convenient. In this case, $C_\ell^{\kappa\kappa} = \ell^2 C_\ell^{dd}/4$.

➲ *Figure 13-29* shows the lensing kernel, the integrand in (➲ 13.15), as a function of the conformal look-back distance η for two different values of ℓ. Most of the dependence on ℓ arises because different values of ℓ pick out different ranges of wavelengths k from the power spectrum $P_\Psi(k = (\ell+1/2)/D(z), z)$. Note that the kernel picks out a broad distribution at relatively high z. This is why lensing is especially sensitive to the sum of neutrino masses and early time dark energy.

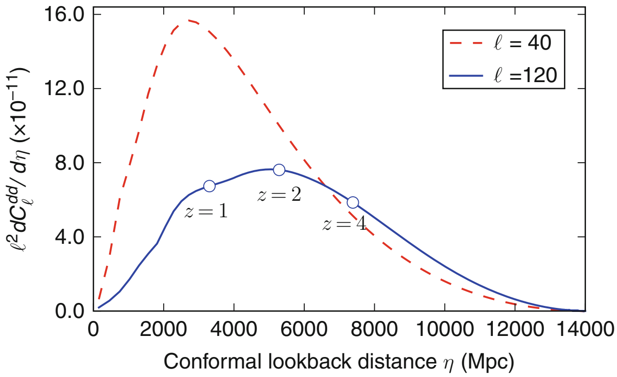


Fig. 13-29

The CMB lensing kernel as a function of conformal look-back distance, η. The decoupling surface is at η_{dec} = 14,000 Mpc for z_{dec} = 1,090. Note that different values of ℓ weight contributions from our past differently (Figure courtesy of Sudeep Das)

For the CMB, the lensing distortion leads to a change in statistical distribution of the fluctuations. Without lensing, the distribution of CMB temperature fluctuations is Gaussian, and the phases are uncorrelated. Lensing correlates the phases, as is so evident in ➋ *Fig. 13-28*, and adds kurtosis (4-pt function) to the Gaussian. The source of the correlations may be seen in (➋ 13.13). An expansion gives $T(\boldsymbol{n}) = T^u(\boldsymbol{n}) + \nabla T^u(\boldsymbol{n})\vec{d}(\boldsymbol{n}) + \ldots$. The second term is significant when the CMB temperature gradients and deflections are large. From ➋ *Figs. 13-21* and ➋ *13-22* we see that the temperature gradients are high on the degree angular scales corresponding to the first peak. These correspond to $\ell \sim 100$. From (➋ 13.14), one finds that the rms deflection is roughly 2.5 arcmin. This angular scale corresponds to $\ell \sim 3{,}000$. CMB lensing is dominated by deflections of 2.5 arcmin that are coherent over degree angular scales. Thus, the power spectrum of $\ell^2 C_\ell^{dd}$ has a broad hump near $\ell = 50$.

To observe lensing, one wants arcmin resolution maps of the CMB that have good statistical properties at degree angular scales. Currently, the Planck, ACT, and SPT maps satisfy these criteria. The connection between the CMB maps and the above is made with an optimal quadratic estimator (Hu and Okamoto 2002; Das et al. 2011a):

$$\begin{aligned}(2\pi)^2\delta(L-L')C_L^{\kappa\kappa} = |N^\kappa(L)|^2 \int \frac{d^2\ell}{(2\pi)^2}\int \frac{d^2\ell'}{(2\pi)^2}|g(\ell,L)|^2 \\ \times\{T^*(\ell)T^*(L-\ell)T(\ell')T(L'-\ell') - < T^*(\ell)T^*(L-\ell)T(\ell') \\ T(L'-\ell') >_{\text{Gauss}}\}, \qquad (13.16)\end{aligned}$$

where ℓ and ℓ' are for the high-resolution CMB maps, g is a filter that can be tuned to optimize signal-to-noise, and N is a normalization. ➋ Equation 13.17 manifestly shows that the convergence power spectrum is related to the four-point function. The second term in brackets is the Gaussian part of the four-point function. This must be subtracted from the full expression to isolate the terms that are responsible for correlating the phases. The Gaussian part can be determined by randomizing the phases in the 2D transform of the map (Dvorkin and Smith 2009; Hanson et al. 2011; Das et al. 2011a).

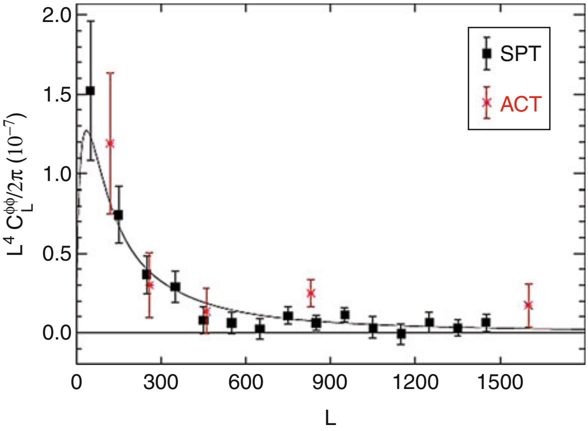


Fig. 13-30
The CMB lensing power spectrum, $C^{\kappa\kappa}$, and measurements from the ACT and SPT teams. The theoretical power spectrum peaks at L ~ 50 (Figure from van Engelen et al. (2012))

Lensing of the CMB was first clearly seen through a cross-correlation with radio galaxy counts (Smith et al. 2007) and the Sloan Digital Sky Survey (Hirata et al. 2008). There was also ~3σ evidence for it from the power spectrum of the CMB (Lueker et al. 2010; Das et al. 2011b). The first detection that was rooted in the characteristic aspects of lensing of the CMB alone was reported in Das et al. (2011a) and based on the ACT data. The lensing power spectrum, $C^{\kappa\kappa}$, was detected at 4σ. Next, van Engelen et al. (2012)) reported a more than 6σ detection of $C^{\kappa\kappa}$ from the SPT. The results are shown in *Figs. 13-30* and *13-31*.

CMB lensing is a new observational handle on the cosmos. The deflection field is a measure of the effects of the matter fluctuations between us and the decoupling surface and thus probes different physical processes than does the primary anisotropy. Thus, with lensing, one uses the CMB to extract information about the volume of the universe and breaks the "geometric degeneracy" associated with the primary anisotropy. This is an exciting direction for observations of the CMB. For example, it means that with only the CMB, we can determine the geometry of the universe or, in other words, using only the CMB, one can deduce that there must be a dark-energy term. Because the signal is intrinsically non-Gaussian, it also means that there is reason to push beyond the cosmic variance limited measurements of the temperature anisotropy.

Figure 13-32, adapted from Smith et al. (2009), shows the sensitivity of the lensing power spectrum to various cosmological parameters of interest. To connect with the measurements we just discussed, we examine curvature. At L ~ 300, the change in $C_L^{\kappa\kappa}$ with Ω_k is −0.4. In other words, to constrain Ω_k to 25%, one must measure $C_L^{\kappa\kappa}$ to 0.1. This is roughly the size of the error bars in *Fig. 13-30*.

A number of studies have been made about how well one could do in principle on the parameters to which lensing is particularly sensitive. A review based on a possible future satellite mission is given in Smith et al. (2009). For an instrument with a resolution of 5 arcmin and a sensitivity of 4 μK-arcmin (1σ noise of 4 μK for each 1 arcmin square pixel) over the full sky, one can reach a 1σ sensitivity on the mass of the neutrino of 0.05 eV, on the equation of state for early time dark energy of 0.15, and on Ω_k of 0.0025. One obtains similar sensitivities by observing a quarter of the sky with 2 arcmin resolution to a depth of 5 μK-arcmin. As a point of reference, at 150 GHz the Planck satellite has a resolution of roughly 7 arcmin and is expected

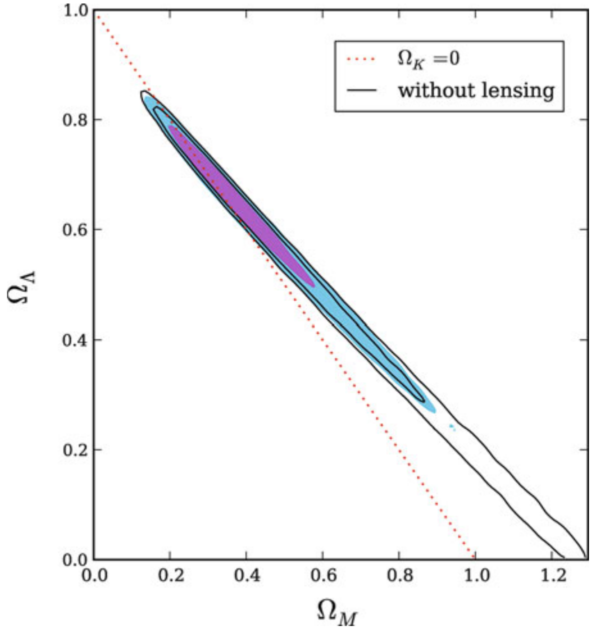


Fig. 13-31

The space of models that fit the CMB alone similar to those shown in *Fig. 13-23*. The *solid black* lines show the 1σ and 2σ contours for what may achieved with the CMB alone without lensing. Note that an $\Omega_\Lambda = 0, \Omega_m = 1.23$ is a perfectly acceptable model. The colored contours show the same constraints but with the ACT lensing included (Sherwin et al. (2011)). The geometric degeneracy is broken and $\Omega_\Lambda = 0$ is excluded at 3.8σ

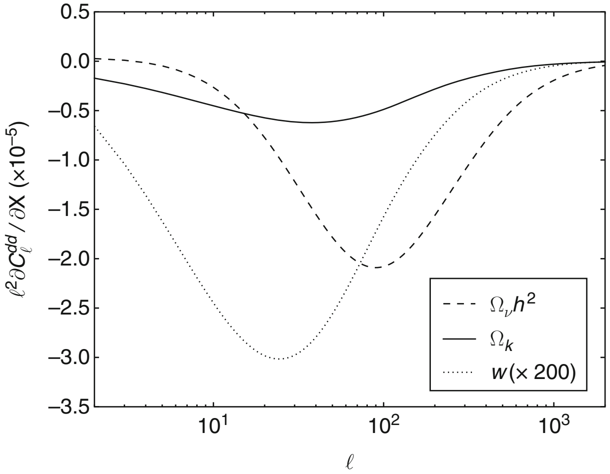


Fig. 13-32

The sensitivity to the lensing power spectrum (*Fig. 13-30*) to the sum of neutrino masses, early dark energy, and spatial curvature, Ω_k (Figure courtesy of Sudeep Das)

to achieve an average of ≈50 μK-arcmin over the whole sky. From the ground, ACT and SPT are achieving ≈20 μK-arcmin with a resolution of 1–1.5 arcmin over significant regions of sky.

The CMB lensing deflection field may also be correlated with the SZ effect, galaxy shear, quasars, the LRGs, and a host of other phenomena to find the growth rate of structure. The growth rate in turn is another probe of dark energy and the mass of the neutrino. CMB lensing may be combined with many other cosmic probes. For example, when combined with the LSST and Planck (e.g., Joudaki and Kaplinghat 2011), one may in principle determine the equation of state of the early dark energy to better than a percent and curvature to 0.06%.

7.2.3 Neutrinos

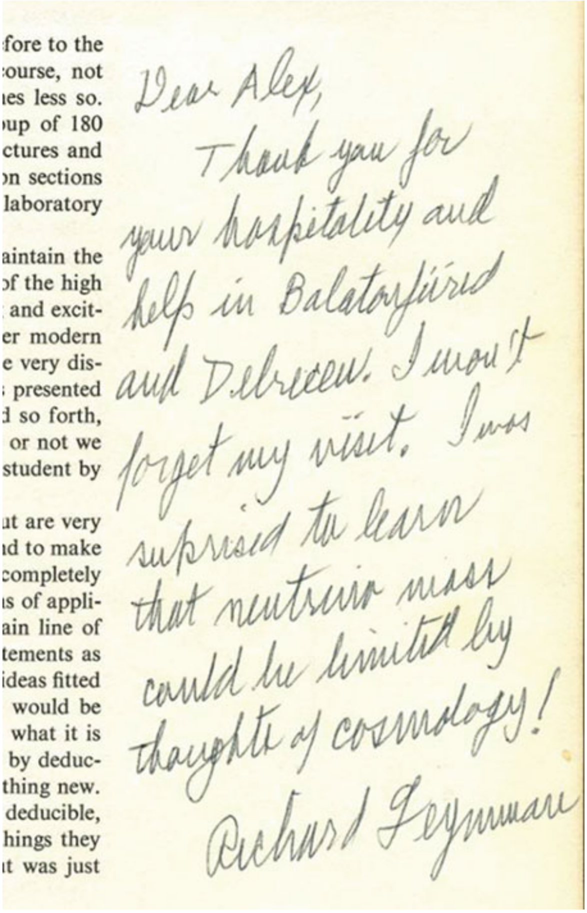
The properties of neutrinos affect the appearance of the CMB. To be more precise, a weakly interacting relativistic particle in the early universe can affect the CMB just as neutrinos would. But because neutrinos exist, we associate them with this primordial constituent. There could well be neutrinos plus additional weakly interacting relativistic particles, but for the purposes of this chapter, we will lump them all together. Neutrinos affect the CMB through their response to gravity. Thus, from the CMB we cannot tell that the particles are, for example, Dirac or Majorana particles (the signature is the same for both) or even if they are spin 1/2. However, we can tell the number of relativistic species and the sum of the neutrino masses. With the technologies currently being developed, we will be able to determine the mass sum in multiple independent ways to a level of roughly 0.06 eV, near the current *lower* limit set by atmospheric neutrino oscillations.

Before getting into details, it is worth noting that neutrinos have been part of the cosmological picture since the earliest days. Understanding them was critical to Big Bang nucleosynthesis. At one point, hints of a 30 eV mass led many to consider them as a major cosmological constituent. In 1972, Alex Szalay was a college student at the Eötvös University in Hungary. Following the advice of Zel'dovich, he computed how one could find the neutrino mass from cosmological observations as part of his undergraduate thesis. That year, the Neutrino '72 conference meet in Balatonfüred, Hungary. Szalay's work on neutrinos was presented at the conference. After the conference, Feynman wrote him the note in ➋ *Fig. 13-33*. His undergraduate work grew into his Ph.D. thesis; computing the effect of neutrinos on density fluctuations would later become Szalay and Marx (1976). The first link between the CMB and neutrino mass was presented in Doroshkevich et al. (1981) and Bond and Szalay (1983).

First consider electron neutrinos. At a redshift of ~10^{10}, electrons,[5] positrons, and photons were strongly coupled through the reaction $\gamma + \gamma \leftrightarrow e^+ + e^-$. In turn, the electrons and positrons were coupled to the neutrinos through $\nu + \bar{\nu} \leftrightarrow e^+ + e^-$. All the particles were highly relativistic. At these early times, the energy density in photons was σT_γ^4 and that of neutrinos was $(7/8)\,\sigma T_\nu^4$ with $T_\nu = T_\gamma$ and σ the Stefan-Boltzmann constant. The factor of 7/8 is from the integral over a Fermi-Dirac distribution as opposed to a Bose-Einstein distribution. When the temperature of the universe cooled to $T \approx m_e c^2/k$, or $z \approx 2 \times 10^9$, the vast majority of electrons and positrons annihilated and dumped their energy into the photons. This increased the temperature of the photons relative to the neutrinos by $T_\gamma = (11/4)^{1/3}\,T_\nu$. Thus, today $T_\nu = 1.95$ K.

The neutrinos today are still distributed according to an orbital occupation number of $n_\nu = \frac{2}{e^{q/kT_\nu}+1}$, where $q = pac$ with p the relativistic momentum, $T_\nu = 1.95$ K, a the scale factor, and

[5] $m_e c^2 \sim 0.5$ MeV corresponds to $T = 6 \times 10^9$ K.



fore to the
ourse, not
es less so.
up of 180
ctures and
n sections
laboratory

aintain the
f the high
and excit-
er modern
e very dis-
presented
l so forth,
or not we
student by

ut are very
d to make
completely
s of appli-
ain line of
tements as
ideas fitted
would be
what it is
by deduc-
thing new.
deducible,
hings they
t was just

Dear Alex,
Thank you for
your hospitality and
help in Balatonfüred
and Debrecen. I won't
forget my visit. I was
surprised to learn
that neutrino mass
could be limited by
thoughts of cosmology!
Richard Feynman

Fig. 13-33

Note from Richard Feynman to Alex Szalay inscribed inside Szalay's copy of the Feynman lectures

the factor of 2 accounts for the two spin states. This occupation number does not change if the neutrinos are massive. In other words, the occupation is set when they are highly relativistic. Because the expansion of the universe is adiabatic, the neutrinos do not change orbitals as the universe evolves.

The energy density today for a single species is given by

$$\rho = \frac{8\pi}{(hc)^3}\frac{1}{a^4}\int_0^\infty q^2 n_\nu \sqrt{q^2 + m_\nu^2 c^2 a^2}\, dq. \qquad (13.17)$$

In Fig. 13-34, the distribution of the neutrino background (the integrand of (13.17)) is compared to that of the Planck distribution for the CMB. If $m_\nu = 0$, then the energy density as a function of momentum, p, is similar to that of the photons, but the peak is shifted to the left because the neutrinos are slightly colder. In addition, the low momentum tail falls off more quickly than for photons due to the "+" sign in the denominator of n_ν as opposed to "−" sign for photons. As the neutrino mass increases, the energy term in (13.17) becomes independent of q, and thus the low momentum tail of the distribution is proportional to q^2. The photons have

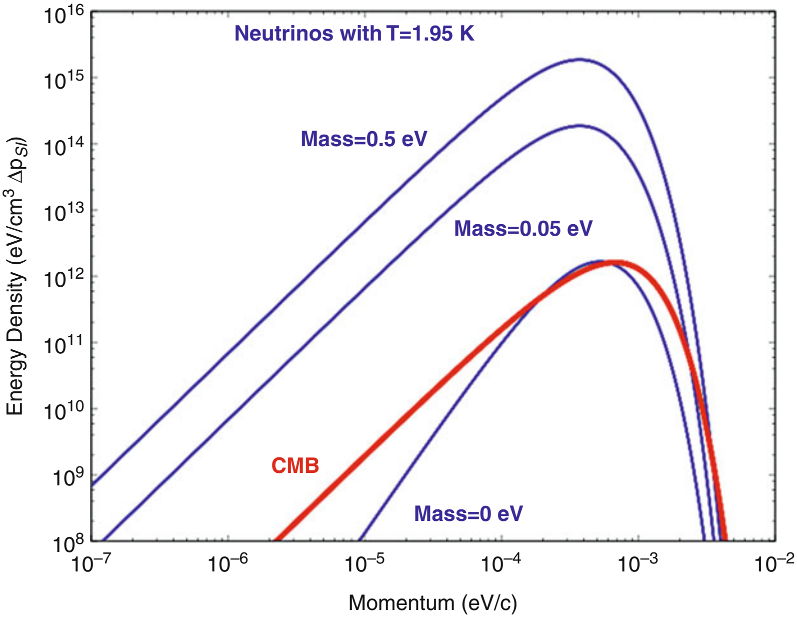


Fig. 13-34

The neutrino energy distribution as a function of neutrino momentum. The integral of the *top curve* tells us that the current energy density for a species of neutrino with a mass of 0.5 eV is 55 eV/cm^3. We know the mass of one species must be at least 0.05 eV, and so for these the energy density is 5.5 eV/cm^3. For the CMB, the energy density is 0.28 eV/cm^3 today

the same distribution at low momentum but for a different reason; here, the q dependence in the denominator of the occupation number cancels the q in the energy dependence.

When thinking about neutrinos, the two most important times in cosmic evolution are matter-radiaton equality and decoupling. In the the standard six-parameter, Λ-dominated flat cosmology, these occur at $z_{\mathrm{eq}} = 3{,}140$ and $z_{\mathrm{dec}} = 1{,}090$ respectively (Komatsu et al. 2011). The essential feature of z_{eq} is that for $z > z_{\mathrm{eq}}$ the expansion rate of the universe slows down and the formation of cosmic structure can begin. Neutrinos, for any mass compatible with the data, act relativistically at matter-radiation equality. By relativistic we mean $kT_\nu \gtrsim m_\nu c^2$. *Figure 13-35* shows the energy density of various cosmic constituents as a function of scale factor of the universe. For neutrinos, blue lines, this is simply ρ in (13.17). The energy density in radiation (the CMB) scales as $1/a^4$, the energy density in matter scales as $1/a^3$, and the energy density in a cosmological constant is independent of a.

From *Fig. 13-35* we can gain an intuition for how neutrinos affect the CMB. First, let us examine the number of relativistic species, N_{eff}. We know there are three families of neutrinos, and so from particle physics, expect $N_{\mathrm{eff}} = 3$. In the cosmological context, we define N_{eff} through

$$\rho_{\mathrm{rad}} \equiv \left(1 + \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} N_{\mathrm{eff}}\right)\rho_\gamma. \tag{13.18}$$

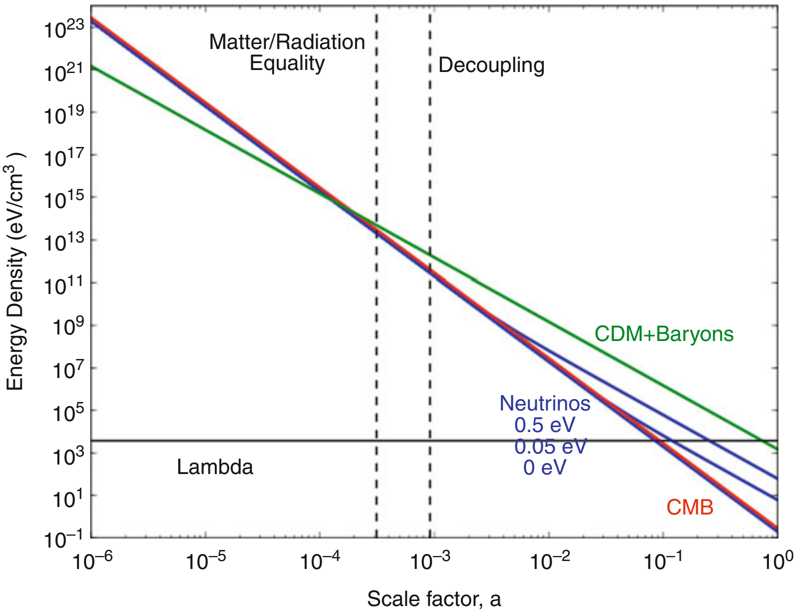


Fig. 13-35
The scaling of the cosmic constituents with scale factor *a*. Green is for CDM and baryons, *blue* is for neutrinos of three different rest masses, 0.5, 0.05, and 0 eV, *red* is the CMB, and *solid black* is the cosmological constant. Note that for all viable neutrino masses, neutrinos scale with the expansion like radiation before $a \approx 2 \times 10^3$ and like matter now. The mass corresponding to $T_\nu = 1.95$ K is $m_\nu = 0.2$ meV

Because the annihilation of e^+ with e^- does not result in all of the energy going into photons alone, some goes to the neutrinos, and thus $N_{\text{eff}} = 3.04$ in (13.18) (Mangano et al. 2005; Kneller and Steigman 2004). If $N_{\text{eff}} > 3.04$, then in *Fig. 13-35*, the blue lines near z_{eq} and z_{dec} are shifted up. As a result, z_{eq} is closer to z_{dec}. This means that prior to decoupling, the higher N_{eff}, the greater the expansion rate. This is a result of the fact that a radiation-dominated universe expands more rapidly than does a matter-dominated one. As pointed out in Bashinsky and Seljak (2004) and Hou et al. (2011), this leads to increased Silk damping and thus a suppression of the anisotropy at $l > 1{,}000$. An outline of the argument is as follows: the diffusion scale of a photon, r_d, scales as $1/\sqrt{H}$, where H is the expansion parameter prior to decoupling. If r_d increases, there is more diffusion and thus more Silk damping. One way to view the scaling (Hou et al. 2011) is that r_d increased as $t^{1/2}$ as one would expect of diffusion. However, as H increases, t decreases and r_d decreases. For the CMB, the other key scale is the acoustic horizon, $r_A = \theta_A D_a$ with D_a the angular diameter distance, and θ_A the angle. The quantity θ_A is precisely determined by the $\ell < 500$ anisotropy data to be $\theta_A = 0.5953 \pm 0.0014^\circ$ (Komatsu et al. 2011). The sound horizon, r_A, is essentially the sound speed multiplied by time and so scales as $1/H$. Since θ_A is so well constrained by the data, the relevant quantity is r_d/r_A which scales as H. Thus, the more relativistic species there are, the larger H at the decoupling era and the larger the damping scale relative to the acoustic scale. The result is more suppression of the $l > 1{,}000$ damping tail relative to the acoustic peaks.

The damping tail is also affected by the fraction of helium. The more helium, the more electrons are bound up in atoms, the longer the diffusion length for a photon, and therefore the greater the suppression of the CMB fluctuations. As a result, there is some degeneracy between the primordial helium fraction and the number of relativistic species. The two effects, though, can be separated (see, e.g., Jungman et al. 1996; Dunkley et al. 2011).

Massive neutrinos affect how we interpret the CMB in a number of ways. For example, for all allowed mass sums ($0.05 < \Sigma_\nu < 0.5$ eV), the neutrinos are nonrelativistic today and should be counted as part of the matter budget as indicated on the right hand side of *Fig. 13-35*. However, for the same mass range, they are part of the radiation budget at decoupling. Thus, to constrain the mass, one compares measures of the mass fluctuations at low redshift, for example, with the σ_8 parameter, to the fluctuations that give rise to the CMB.

The mass of the neutrino also directly affects the acoustic peak structure and the growth rate of structure. There are multiple effects at play as discussed in, for example, Dodelson et al. (1996) and Ichikawa et al. (2005) (see also Hannestad and Brandbyge 2010). The driving concepts are that at $z_{\rm dec}$ the neutrino temperature is roughly 2,000 K corresponding to 0.1 eV, and so massive neutrinos are in the process of becoming nonrelativistic. Also, while photons in the decoupling epoch diffuse out of potential wells as they scatter off electrons, neutrinos free stream out of potential wells as their interactions are minimal. The net effect is that the phase and amplitudes of the acoustic peaks are slightly altered and that the more massive the neutrino, the greater the suppression of the formation of cosmic structure at small scales.

This suppression of structure leads to a decrease in the CMB lensing signal as evident in *Fig. 13-32*. Note that this signal is based solely on CMB observations and does not depend on additional measures of galaxy spectra or σ_8. As noted above, the signal can be extracted from the non-Gaussianity of the CMB. Also note how characteristic the signal is in the lensing spectrum. There is a second way to see the neutrino mass signature that offers a built-in cross check. At high ℓ, B-mode polarization is produced by the gravitational lensing of the E-modes (Zaldarriaga and Seljak 1998). This is the same mechanism that leads to a confounding astrophysical signal at low ℓ that hides the primordial B-modes. It is convenient that in polarization this region is relatively free of foreground emission. *Figure 13-36* shows the effect. If the neutrino mass sum is 0.5 eV, the lensing B-mode signal is suppressed by roughly 25%.

7.3 The Future

There is still much to be learned from the CMB. The advantages of observing from space are so strong that we should anticipate a future satellite mission. Space offers long uninterrupted periods of observation from a platform whose thermal stability can be measured in milliKelvin. This combination enables a myriad of possible consistency checks for various systematic effects. The history of the field is one of great advances made with ground-based and suborbital experiments that then inform a satellite design. Detector sensitivities near 2 mm wavelength are approaching levels where they will be limited by the photon noise from the CMB itself. Soon this will be the case across the frequency band where the CMB dominates. We have learned how to produce polarization-sensitive arrays of hundreds to thousands of detectors that are read out with low-power superconducting electronics. There are no known impediments to even larger arrays. Many groups are investigating new kinds of radiometers based on technologies ranging from multimoded detectors and optics to multichroic detectors with broad-band optics. There are advances with both bolometric and coherent systems. The field is dynamic. The instruments

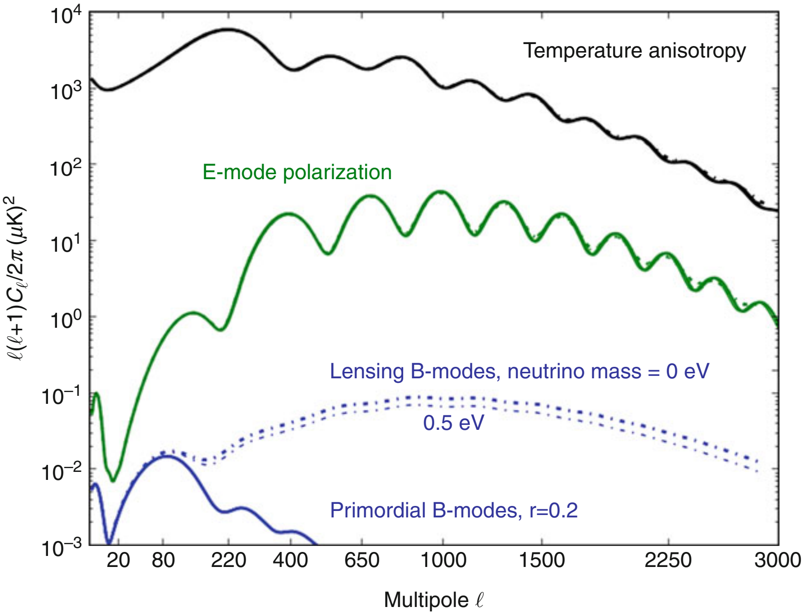


Fig. 13-36
The effect of massive neutrinos on the CMB polarization spectrum. Matter fluctuations between us and the surface of last scattering lens the CMB. This can be detected directly in at least two ways. One can isolate the non-Gaussian aspect part of the temperature anisotropy and extract the B-mode signal from that. Another way is to measure B-modes that result from lensing of the E-modes. This is a particularly clean signal near these wavelengths. This figure shows the difference in the B-mode signal for neutrinos with a 0 mass sum and those with an 0.5 eV mass sum. For comparison, the primordial B-mode signal with $r = 0.2$ is also shown. Current error bars on the temperature anisotropy are near 1 μK^2, and so measuring the high-ℓ B-modes is not far off

being developed define the forefront in receiver technology. We expect many more exciting results over the coming years. However, as we probe ever deeper to search for more subtle aspects of nature, a future space mission will be required.

Acknowledgments

It is a pleasure to acknowledge the work of our many colleagues who we have cited in this work. Paul Richards, Rainer Weiss, and David Wilkinson were pioneers who showed many of us the way to think and measure the CMB. Among living pioneers, Rashid Sunyaev and Jim Peebles, among many others, worked out how the universe could or should be, long before we were able to know that they were right. Pat Thaddeus initiated work on the COBE mission proposal, and Nancy Boggess at NASA Headquarters protected it from those who saw it as only a physics experiment. Jens Chluba and Dale Fixsen read the spectrum portion of our manuscript and made many useful comments. We thank Sudeep Das for useful comments and for providing figures on CMB lensing. Shaul Hanany, Colin Hill, Bill Jones, Akito Kusaka, Avi Loeb, and Suzanne Staggs made a number of helpful comments.

References

Addison, G. E., et al. 2012, ApJ, 752(2), article id. 120
Aguirre, A. 1999, ApJ, 521, 17
Aguirre, A. N. 2000, ApJ, 533, 1
Albrecht, A., & Steinhardt, P. J. 1982, Phys. Rev. Lett., 48, 1220
Ali-Haïmoud, Y., & Hirata, C. M. 2011, Phys. Rev. D, 83, 043513
Ali-Haïmoud, Y., Hirata, C. M., & Dickinson, C. 2009, MNRAS, 395, 1055
Alizadeh, E., & Hirata, C. M. 2011, Phys. Rev. D, 84, 083011
Alpher, R. A., & Herman, R. C. 1948, Phys. Rev., 74, 1737
Bardeen, J. M., Steinhardt, P. J., & Turner, M. S. 1983, Phys. Rev. D, 28, 679
Bashinsky, S., & Seljak, U. 2004, Phys. Rev. D, 69, 083002
Basko, M. M., & Polnarev, A. G. 1980, MNRAS, 191, 207
Baumann, D., et al. 2009, in American Institute of Physics Conf. Ser. 1141, ed. S. Dodelson, D. Baumann, A. Cooray, J. Dunkley, A. Fraisse, M. G. Jackson, A. Kogut, L. Krauss, M. Zaldarriaga, & K. Smith, 10–120
Benford, D. J., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., 7741
Bennett, C. L., et al. 1996, ApJL, 464, L1
Bennett, C. L., et al. 2003a, ApJS, 148, 97
Bennett, C. L. et al. 2003b, ApJ, 583, 1
Benoît, A., et al. 2003, A&A, 399, L19
Birkinshaw, M., Gull, S. F., & Hardebeck, H. 1984, Phy. Rev. Lett., 309, 34
Bischoff, C., et al. 2008, ApJ, 684, 771
Blanchard, A., & Schneider, J. 1987, A&A, 184, 1
Bock, J., et al. 2006, ArXiv Astrophysics e-prints
Boggess, N. W., et al. 1992, ApJ, 397, 420
Bond, J. R. 1996, in Cosmology and Large Scale Structure, Les Houches Session LX, ed. R. Schaeffer, J. Silk, M. Spiro, & J. Zinn-Justin (London: Elsevier), 469
Bond, J. R., & Szalay, A. S. 1983, ApJ, 274, 443
Bondi, H., & Gold, T. 1948, MNRAS, 108, 252
Bracewell, R. N., & Conklin, E. K. 1968, Nature, 219, 1343
Brown, M. L., et al. 2009, ApJ, 705, 978
Burigana, C., et al. 1991, ApJ, 379, 1
Burns, J. O., et al. 2012, Adv. Space Res., 49, 433
Carbone, C., Baccigalupi, C., Bartelmann, M., Matarrese, S., & Springel, V. 2009, MNRAS, 396, 668
Carlstrom, J. E., et al. 2011, PASP, 123, 568
Charlassier, R., for the BRAIN Collaboration 2008, in Proceeding of the 43rd "Rencontres de Moriond" on Cosmology, La Thuile, Italy, March 15–22, 2008
Chiang, H. C., et al. 2010, ApJ, 711, 1123
Chluba, J., Erickcek, A. L., & Ben-Dayan, I. 2012a, ArXiv e-prints
Chluba, J., Khatri, R., & Sunyaev, R. A. 2012b, ArXiv e-prints
Chluba, J., & Sunyaev, R. A. 2006, A&A, 458, L29
Chluba, J., & Sunyaev, R. A. 2012a, MNRAS, 419, 1294
Chluba, J., & Sunyaev, R. A. 2012b, MNRAS, 419, 1294
Conklin, E. K. 1969, Nature, 222, 971
Coppola, C. M., D'Introno, R., Galli, D., Tennyson, J., & Longo, S. 2012a, ApJS, 199, 16
Coppola, C. M., D'Introno, R., Galli, D., Tennyson, J., & Longo, S. 2012b, ApJS, 199, 16
Corey, B. E., & Wilkinson, D. T. 1976, in Bulletin of the American Astronomical Society, Vol. 8, 351
Crill, B. P., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., 7010
Daly, R. A. 1991, ApJ, 371, 14
Das, S., et al. 2011a, Phys. Rev. Lett., 107, 021301
Das, S., et al. 2011b, ApJ, 729, 62
de Bernardis, P., et al. 2000, Nature, 404, 955
de Oliveira-Costa, A., Kogut, A., Devlin, M. J., Netterfield, C. B., Page, L. A., & Wollack, E. J. 1997, ApJL, 482, L17
de Vega, H. J., & Sanchez, N. G. 2010, MNRAS, 404, 885
Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. 1965, ApJ, 142, 414
Dodelson, S., Gates, E., & Stebbins, A. 1996, ApJ, 467, 10
Doroshkevich, A. G., Khlopov, M. I., Sunyaev, R. A., Szalay, A. S., & Zeldovich, I. B. 1981, Ann. NY Acad. Sci., 375, 32
Draine, B. T., & Lazarian, A. 1998, ApJ, 508, 157
Draine, B. T., & Lazarian, A. 1999, ApJ, 512, 740
Dubrovich, V. K. 1975, Sov. Astron. Lett., 1, 196

Dubrovich, V. K., & Stolyarov, V. A. 1995, A&A, 302, 635

Dunkley, J., et al. 2009, in American Institute of Physics Conf. Ser., 1141, ed. S. Dodelson, D. Baumann, A. Cooray, J. Dunkley, A. Fraisse, M. G. Jackson, A. Kogut, L. Krauss, M. Zaldarriaga, & K. Smith, 222–264

Dunkley, J., et al. 2011, ApJ, 739, 52

Dvorkin, C., & Smith, K. M. 2009, Phys. Rev. D, 79, 043003

Eimer, J. R., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., 7733

Einstein, A. 1917, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin), Seite 142–152., 142

Essinger-Hileman, T., et al. 2010, in Astrophysics - Instrumentation and Methods for Astrophysics, Astrophysics - Cosmology and Extragalactic Astrophysics. Proceedings of the Thirteenth International Conference on Low-Temperature Detectors

Feng, J. L., Rajaraman, A., & Takayama, F. 2003, Phys. Rev. D, 68, 063504

Fixsen, D. J. 2009, ApJ, 707, 916

Fixsen, D. J., & Dwek, E. 2002, ApJ, 578, 1009

Fixsen, D. J., Cheng, E. S., Gales, J. M., Mather, J. C., Shafer, R. A., & Wright, E. L. 1996, ApJ, 473, 576

Fixsen, D. J., Hinshaw, G., Bennett, C. L., & Mather, J. C. 1997, ApJ, 486, 623

Fixsen, D. J., Bennett, C. L., & Mather, J. C. 1999, ApJ, 526, 207

Fixsen, D. J., Wollack, E. J., Kogut, A., Limon, M., Mirel, P., Singal, J., & Fixsen, S. M. 2006, Rev. Sci. Instrum., 77, 064905

Fowler, J. W., et al. 2007, Appl. Opt., 46, 3444

Friedman, A. 1922, Z. Phys., 10, 377

Fukugita, M., & Kawasaki, M. 1990, ApJ, 353, 384

Galli, S., Bean, R., Melchiorri, A., & Silk, J. 2008, Phys. Rev. D, 78, 063532

Ganga, K., Cheng, E., Meyer, S., & Page, L. 1993, ApJL, 410, L57

Gervasi, M., Zannoni, M., Tartari, A., Boella, G., & Sironi, G. 2008, ApJ, 688, 24

Gnedin, N. I., & Ostriker, J. P. 1992, ApJ, 400, 1

Gold, B., et al. 2011, ApJS, 192, 15

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Guth, A. H., & Pi, S.-Y. 1982, Phys. Rev. Lett., 49, 1110

Gush, H. P., Halpern, M., & Wishnow, E. H. 1990, Phys. Rev. Lett., 65, 537

Hall, N. R., et al. 2010, ApJ, 718, 632

Halverson, N. W., et al. 2002, ApJ, 568, 38

Hanany, S., et al. 2000, ApJL, 545, L5

Hanany, S., Niemack, M., & Page, L. 2012, CMB Optics (Springer), Planets, Stars and Stellar Systems (PSSS)

Hand, N., et al. 2012, ArXiv e-prints

Hannestad, S., & Brandbyge, J. 2010, JCAP, 3, 20

Hanson, D., Challinor, A., Efstathiou, G., & Bielewicz, P. 2011, Phys. Rev. D, 83, 043005

Harari, D. D., & Zaldarriaga, M. 1993, Phys. Lett. B, 319, 96

Harrison, E. R. 1970, Phys. Rev. D, 1, 2726

Haslam, C. G. T., Klein, U., Salter, C. J., Stoffel, H., Wilson, W. E., Cleary, M. N., Cooke, D. J., & Thomasson, P. 1981, A&A, 100, 209

Hawking, S. W. 1982, Phys. Lett. B, 115, 295

Herzberg, G. 1950, Molecular Spectra and Molecular Structure. Vol.1: Spectra of Diatomic Molecules (New York: Van Nostrand Reinhold)

Hincks, A. D., et al. 2010, ApJS, 191, 423

Hinshaw, G., et al. 2003, ApJS, 148, 135

Hinshaw, G., et al. 2009, ApJS, 180, 225

Hirata, C. M., Ho, S., Padmanabhan, N., Seljak, U., & Bahcall, N. A. 2008, Phys. Rev. D, 78, 043520

Hooper, D., & Linden, T. 2011, Phys. Rev. D, 83, 083517

Hou, Z., Keisler, R., Knox, L., Millea, M., & Reichardt, C. 2011, ArXiv e-prints

Howlett, C., Lewis, A., Hall, A., & Challinor, A. 2012, JCAP, 4, 27

Hoyle, F. 1948, MNRAS, 108, 372

Hu, W., & Okamoto, T. 2002, ApJ, 574, 566

Hu, W., Scott, D., & Silk, J. 1994, ApJL, 430, L5

Hubble, E. 1929, Proc. Natl. Acad. Sci., 15, 168

Ichikawa, K., Fukugita, M., & Kawasaki, M. 2005, Phys. Rev. D, 71, 043001

Illarionov, A. F., & Sunyaev, R. A. 1974, Astronomicheskii Zh., 51, 1162

Illarionov, A. F., & Sunyaev, R. A. 1975a, Sov. Astron., 18, 413

Illarionov, A. F., & Sunyaev, R. A. 1975b, Sov. Astron., 18, 691

Jarosik, N., et al. 2011, ApJS, 192, 14

Jedamzik, K., Katalinić, V., & Olinto, A. V. 2000, Phys. Rev. Lett., 85, 700

Joudaki, S., & Kaplinghat, M. 2011, ArXiv e-prints

Jungman, G., Kamionkowski, M., Kosowsky, A., & Spergel, D. N. 1996, Phys. Rev. D, 54, 1332

Kamionkowski, M., Kosowsky, A., & Stebbins, A. 1997, Phys. Rev. D, 55, 7368

Katayama, N., & Komatsu, E. 2011, ApJ, 737, 78

Keisler, R., et al. 2011, ApJ, 743(1), article id. 28

Kelsall, T., et al. 1998, ApJ, 508, 44

Khatri, R., Sunyaev, R. A., & Chluba, J. 2011, arXiv:1110.0475

Kneller, J. P., & Steigman, G. 2004, New J. Phys., 6, 117

Khoury, J., Steinhardt, P. J., & Turok, N. 2003, Phy. Rev. Lett., 91, 16, 161301

Knox, L., & Page, L. 2000, Phys. Rev. Lett., 85, 1366

Kogut, A., Banday, A. J., Bennett, C. L., Górski, K. M., Hinshaw, G., & Reach, W. T. 1996, ApJ, 460, 1

Kogut, A., Wollack, E., Fixsen, D. J., Limon, M., Mirel, P., Levin, S., Seiffert, M., & Lubin, P. M. 2004, Rev. Sci. Instrum., 75, 5079

Kogut, A., et al. 2011, J. Cosmol. Astropart. Phys., 7, 25

Komatsu, E., et al. 2009, ApJS, 180, 330

Komatsu, E., et al. 2011, ApJS, 192, 18

Kosowsky, A., & Turner, M. S. 1995, Phys. Rev. D, 52, 1739

Kovac, J. M., Leitch, E. M., Pryke, C., Carlstrom, J. E., Halverson, N. W., & Holzapfel, W. L. 2002, Nature, 420, 772

Kuo, C. L., et al. 2004, ApJ, 600, 32

Larson, D., et al. 2011, ApJS, 192, 16

Layzer, D., & Hively, R. 1973, ApJ, 179, 361

Lee, A. T. et al. 2001, ApJL, 561, L1

Leitch, E. M., Readhead, A. C. S., Pearson, T. J., & Myers, S. T. 1997, ApJL, 486, L23

Lemaître, G. 1927, Ann. Soc. Sci. Brux., 47, 49

Lemaître, G. 1931, MNRAS, 91, 490

Lewis, A., & Challinor, A. 2006, Phys. Rep., 429, 1

Linde, A. D. 1982, Phys. Lett. B, 108, 389

Lueker, M., et al. 2010, ApJ, 719, 1045

Maeda, K., Alvarez, H., Aparici, J., May, J., & Reich, P. 1999, A&AS, 140, 145

Mangano, G., Miele, G., Pastor, S., Pinto, T., Pisanti, O., & Serpico, P. D. 2005, Nucl. Phys. B, 729, 221

Martin, D., & Puplett, E. 1970, Infrared Phys., 10, 105

Mather, J. C., et al. 1990, ApJL, 354, L37

Mather, J. C., Fixsen, D. J., Shafer, R. A., Mosier, C., & Wilkinson, D. T. 1999, ApJ, 512, 511

Matsumoto, T., Hayakawa, S., Matsuo, H., Murakami, H., Sato, S., Lange, A. E., & Richards, P. L. 1988, ApJ, 329, 567

Mauskopf, P. D., et al. 2000, ApJL, 536, L59

McDonald, P., Scherrer, R. J., & Walker, T. P. 2001, Phys. Rev. D, 63, 023001

McKellar, A. 1941, Publ. Dominion Astrophys. Obs. Vic., 7, 251

Miller, A. D., et al. 1999, ApJL, 524, L1

Miller, A. D. et al. 2002, ApJS, 140, 115

Montroy, T. E., et al. 2006, ApJ, 647, 813

Mukhanov, V. F., & Chibisov, G. V. 1981, Sov. J. Exp. Theor. Phys. Lett., 33, 532

Netterfield, C. B., Devlin, M. J., Jarosik, N., Page, L., & Wollack, E. J. 1997, ApJ, 474, 47

Netterfield, C. B. et al. 2002, ApJ, 571, 604

Ostriker, J. P., & Cowie, L. L. 1981, ApJL, 243, L127

Ostriker, J. P., & Thompson, C. 1987, ApJL, 323, L97

Page, L., et al. 2007, ApJS, 170, 335

Pardo, J. R., Cernicharo, J., & Serabyn, E. 2001, IEEE Trans. Antennas Propag., 49, 1683

Pearson, T. J., et al. 2003, ApJ, 591, 556

Peebles, P. J. E. 1968a, ApJ, 153, 1

Peebles, P. J. E. 1968b, ApJ, 153, 1

Peebles, P. J., & Wilkinson, D. T. 1968, Phys. Rev., 174, 2168

Peebles, P. J. E., & Yu, J. T. 1970, ApJ, 162, 815

Peebles, P. J. E., Seager, S., & Hu, W. 2000, ApJL, 539, L1

Peebles, P. J. E., Page, L. A., & Partridge, R. B., (ed.) 2009, Finding the Big Bang (Cambridge University Press)

Penzias, A. A., & Wilson, R. W. 1965, ApJ, 142, 419

Planck Collaboration, et al. 2011a, A&A, 536, A1

Planck Collaboration, et al. 2011b, A&A, 536, A8

QUIET Collaboration, et al. 2011, ApJ, 741, 111

Readhead, A. C. S., et al. 2004, ApJ, 609, 498

Rees, M. J. 1968, ApJL, 153, L1

Reich, P., & Reich, W. 1986, A&AS, 63, 205

Reichborn-Kjennerud, B., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., 7741

Riess, A. G., et al. 2011, ApJ, 730, 119

Roger, R. S., Costain, C. H., Landecker, T. L., & Swerdlyk, C. M. 1999, A&AS, 137, 7

Roll, P. G., & Wilkinson, D. T. 1966, Phys. Rev. Lett., 16, 405

Rubiño-Martín, J. A., Chluba, J., & Sunyaev, R. A. 2008, A&A, 485, 377

Ruhl, J. E., et al. 2003, ApJ, 599, 786

Ruhl, J., et al. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser. 5498, ed. C. M. Bradford, P. A. R. Ade, J. E. Aguirre, J. J. Bock, M. Dragovan, L. Duband, L. Earle, J. Glenn, H. Matsuhara, B. J. Naylor, H. T. Nguyen, M. Yun, & J. Zmuidzinas, 11–29

Sachs, R. K., & Wolfe, A. M. 1967, ApJ, 147, 73

Sato, K. 1981, MNRAS, 195, 467

Scott, D., & White, M. 1999, A&A, 346, 1

Seager, S., Sasselov, D. D., & Scott, D. 1999, ApJL, 523, L1

Seager, S., Sasselov, D. D., & Scott, D. 2011, in Astrophysics Source Code Library, record ascl:1106.026, 6026

Seiffert, M., et al. 2011, ApJ, 734, 6

Seljak, U. 1996, ApJ, 463, 1

Sheehy, C. D., et al. 2010, in SPIE Proceedings for Millimeter, Submillimeter and Far-Infrared Detectors and Instrumentation for Astronomy V (Conference 7741, San Diego, CA, USA)

Sherwin, B. D., et al. 2011, Phys. Rev. Lett., 107, 021302

Shirokoff, E., et al. 2011, ApJ, 736, 61

Sievers, J. L., et al. 2007, ApJ, 660, 976

Silk, J. 1968, ApJ, 151, 459

Silk, J., & Stebbins, A. 1983, ApJ, 269, 1

Singal, J., et al. 2011, ApJ, 730, 138

Smith, K. M., Zahn, O., & Doré, O. 2007, Phys. Rev. D, 76, 043510

Smith, K. M., et al. 2009, in American Institute of Physics Conf. Ser. 1141, ed. S. Dodelson, D. Baumann, A. Cooray, J. Dunkley, A. Fraisse, M. G. Jackson, A. Kogut, L. Krauss, M. Zaldarriaga, & K. Smith , 121–178

Smoot, G. F., Gorenstein, M. V., & Muller, R. A. 1977, Phys. Rev. Lett., 39, 898

Smoot, G. F., et al. 1992, ApJL, 396, L1

Sollom, I., Challinor, A., & Hobson, M. P. 2009, Phys. Rev. D, 79, 123521

Staniszewski, Z., et al. 2009, ApJ, 701, 32

Starobinsky, A. A. 1982, Phys. Lett. B, 117, 175

Sunyaev, R. A., & Chluba, J. 2009, Astron. Nachr., 330, 657

Sunyaev, R. A., & Zel'dovich, Y. B. 1970a, Astrophys. Space Sci., 7, 3

Sunyaev, R. A., & Zel'dovich, Y. B. 1970b, Astrophys. Space Sci., 7, 20

Swetz, D. S., et al. 2011, ApJS, 194, 41

Switzer, E. R., & Hirata, C. M. 2005, Phys. Rev. D, 72, 083002

Switzer, E. R., & Hirata, C. M. 2008, Phys. Rev. D, 77, 083006

Szalay, A. S., & Marx, G. 1976, A&A, 49, 437

Taburet, N., Hernández-Monteagudo, C., Aghanim, N., Douspis, M., & Sunyaev, R. A. 2011, MNRAS, 418, 2207

Tashiro, H., Sabancilar, E., & Vachaspati, T. 2012, ArXiv:1202.2474

Tegmark, M. 1996, ApJL, 464, L35

The Polarbear Collaboration, et al. 2010, POLARBEAR the web, arXiv e-print (arXiv:1011.0763)

The QUBIC collaboration, et al. 2011, AP, 34(9), 705–716

Timbie, P. T., et al. 2006, New Astron. Rev., 50, 999

van Engelen, A., et al. 2012, ArXiv e-prints

Varshalovich, D. A., & Khersonskii, V. K. 1977, Sov. Astron. Lett., 3, 155

Vernstrom, T., Scott, D., & Wall, J. V. 2011, MNRAS, 415, 3641

Viero, M. P., et al. 2009, ApJ, 707, 1766

Webster, A. S. 1974, MNRAS, 166, 355

Weiss, R. 1980, ARAA, 18, 489

Partridge, R. B., & Wilkinson, D. T. 1967, Phy. Rev. Lett., 18, 557–559

Winston, R. 1970, J. Opt. Soc. Am. (1917–1983), 60, 245

Woody, D. P., & Richards, P. L. 1979, Phys. Rev. Lett., 42, 925

Woody, D. P., Mather, J. C., Nishioka, N. S., & Richards, P. L. 1975, Phys. Rev. Lett., 34, 1036

Wright, E. L. 1979, ApJ, 232, 348

Wright, E. L. 1982, ApJ, 255, 401

Wright, E. L. 2012, Wright Cosmology Tutorial. http://www.astro.ucla.edu/~wright/CMB.html

Wright, E. L., et al. 1991, ApJ, 381, 200

Wright, E. L., et al. 1992, ApJL, 396, L13

Wright, E. L., et al. 1994, ApJ, 420, 450

Zaldarriaga, M., & Seljak, U. 1997, Phys. Rev. D, 55, 1830

Zaldarriaga, M., & Seljak, U. 1998, Phys. Rev. D, 58, 023003

Zannoni, M., Tartari, A., Gervasi, M., Boella, G., Sironi, G., De Lucia, A., Passerini, A., & Cavaliere, F. 2008, ApJ, 688, 12

Zel'dovich, Y. B. 1963, Sov. J. Exp. Theor. Phys., 16, 1395

Zel'dovich, Y. B. 1972, MNRAS, 160, 1P

Zel'dovich, Y. B., Kurt, V. G., & Sunyaev, R. A. 1969, Sov. J. Exp. Theor. Phys., 28, 146

Index

A

Abell 963, 257
Abell 1367, 210, 225, 226, 235, 236, 242, 256
Abell 2192, 257
Abell 3627, 235, 241, 254
Absolute radiometer for cosmology, astrophysics, and diffuse emission (ARCADE), 612, 627, 636–639
Accretion, 309, 314, 321, 326, 335–348, 350, 365, 368–370, 379, 380, 504–507, 515, 516, 523, 525, 526, 531–535, 540–542, 548, 549, 551–553, 555–557, 559, 561–563
Active galactic nuclei (AGN), 306–381, 504–564
  clustering, 507, 528–530
  feedback, 507, 508, 539–559, 563
  luminosity function, 519, 523, 525, 526, 530, 531, 534
  number counts, 513
ALMA, 257
Anemic galaxies, 220, 226
Anemic spiral galaxies, 30
Angular clustering, 393–394
Anisotropy, 611, 612, 616–621, 624, 625, 627, 629, 630, 634, 635, 640, 644–653, 655–659, 662–680
Annealing, 217
ARCADE. *See* Absolute radiometer for cosmology, astrophysics, and diffuse emission (ARCADE)
Atomic and molecular gas galaxy morphology, 20
Atomic hydrogen, 219–221

B

Bar ansae, 32, 33
Barred galaxies, 6, 17, 18, 23, 24, 27, 32, 33, 39, 42, 49, 66, 80–82
Baryon content, 288–289
BCD galaxies. *See* Blue compact dwarf (BCD) galaxies
BCMs. *See* Brightest cluster members (BCMs)
Bias (in galaxy formation), 397–402, 404–408, 411–414, 417, 418
Black hole formation, 504, 559
Black hole mass function, 505, 533, 540
Blazars, 321, 322, 327, 330–335, 371, 375, 376
Blue compact dwarf (BCD) galaxies, 73–74
Boxy ellipticals, 11
Brightest cluster members (BCMs), 52–53
Butcher–Oemler effect, 210, 211, 255–256, 280, 284

C

Cancer clusters, 225
CDM. *See* Cold dark matter (CDM)
Cepheid, 424–427, 430, 431, 433–446
  distance scale, 434–435, 446
  galactic, 430, 435, 445, 446
CGCG 97-073, 242
CGCG 97-079, 242
Chemical evolution, 477
C IV, 592–595, 598, 599
Classical bulges, 37, 38, 68, 78
Cluster-cluster merger, 257
Cluster-group merger, 254
Clustering, 388–414, 417, 419, 420
CMB. *See* Cosmic microwave background (CMB)
CMD method, 431–432
COBE. *See* Cosmic background explore (COBE)
Cold dark matter (CDM), 618, 645, 660, 667
Collisional ring galaxies, 40, 41, 81, 82
Color galaxy morphology, 71–79
Color-magnitude relation, 273–274
Coma, 210, 220, 223–226, 228, 235, 236, 239–240, 245, 248, 254–257
  cluster, 210, 220, 223, 224, 226, 235, 236, 239–240, 245, 248, 254, 256
Compact elliptical, 230
Content of clusters, 272, 289, 295, 297
Conversion factor, 222
Cooling flows, 269, 271, 290–293, 295
Cooling instability, 286, 291–295
Core-Sérsic model, 94, 98, 100, 102–106
Cosmic background explore (COBE), 612, 615, 616, 618, 623, 625–628, 630–635, 638, 643–648
Cosmic microwave background (CMB), 611–630, 634–636, 639, 640, 642–652, 654–680
  radiation, 612, 644
Cosmic ray gas, 224, 251, 258
Cosmic X-ray background, 515
Cosmology, 453–498
Counter-winding spiral galaxies, 28, 29, 82

D

DA. *See* Dark ages (DA)
Damped Ly α systems, 585–589
Dark ages (DA), 568–570
Dark matter, 388, 390, 397–401, 405–413, 416–420
  halos, 94, 101–102, 131, 399, 401, 406–412, 416–420

Decoupling, 575
Density discontinuity, 237
Density-morphology relation, 271–273
Differential microwave radiometers (DMR), 630, 644–648
Disk galaxies, 213, 216–218, 239, 242
Disky ellipticals, 11, 12
Distance indicators, 424–428, 430, 433, 434
Dust, 184, 192, 223–225, 236, 251, 252, 258
  attenuation, 158, 159
  lanes, 6, 35, 36, 41, 46, 64
  removal, 224
Dust-lane elliptical galaxies, 46
Dwarf elliptical, 211, 223, 229, 230
Dwarf galaxies, 30, 57, 72–76, 211, 216, 229, 230, 237, 245–247
  spheroidal, 74–75
  spiral, 75–76, 82
Dynamical friction, 276, 277, 285

E

Early-type galaxies, 8, 9, 13, 15, 20, 21, 38, 42, 60, 63, 71, 77, 78, 80–83
Eccentric orbits, 213, 226, 254, 256, 257
Eddington luminosity, 520, 534
Einasto model, 102
Elliptical galaxy(ies), 8, 10–14, 36, 37, 42, 46, 47, 66, 69, 77, 83, 216, 226, 235–238, 251, 257
Emission line regions, 347–356
Entropy, 289–290, 300
Environment, 209–258, 271, 272, 276, 281, 284, 285, 302
Environmental effects, 223, 228, 230, 232–247, 258
ESO 137-001, 241, 254
ESO 137-002, 241
Eulerian hydrodynamics, 215, 217, 218, 254
Expanding photosphere method (EPM), 434
Extragalactic distance scale, 424, 426, 435, 439, 445–446

F

Far infrared absoulte spectrophotometer (FIRAS), 612, 615, 616, 620, 624–628, 630–636, 638–640, 642, 643
Feedback, 465, 468, 471–476, 482, 486, 490–492, 494, 495, 497
Filament, 388, 389, 391, 413–419
FIRAS. *See* Far infrared absoulte spectrophotometer (FIRAS)
Fireballs, 245, 247
FIR radio correlation, 231
Flocculent, 6, 27, 28, 64
Formation of clusters, 271, 279, 299–301
Fornax, 210, 230, 235–238, 256
  cluster, 210, 235–238
Fundamental plane, 275–276

G

GAIA. *See* Global Astrometric Interferometer for Astrophysics (GAIA)
Galactic emission, 615, 629, 639, 643, 649, 650, 654
Galatic fountain, 188, 190, 191, 199, 201
Galaxy, 184–202, 209–258
  bars, 130
  bulge-disk ratios, 121–122
  bulgeless, 129–130
  cD, halos, 108
  central mass deficits, 105
  clusters, 209–211, 217, 224–226, 229–231, 235–247, 255–257
  collisions, 272, 276–277
  compact, 94, 95, 118
  dust, 94, 121–124
  elliptical, 97–105, 110, 111, 113, 117, 118, 125, 131
  evolution, 209–258, 391, 405, 406, 411, 413, 415, 416, 418, 482
  excess nuclear light, 94, 105–107
  formation, 454–456, 462–498
  harassment, 212, 213, 229, 230, 256, 257
  morphology, 272, 273
  nuclei, 105
  pseudobulges, 124–129
  scaling relations, 93, 94, 97, 100, 101, 109–120, 128–129, 131
  spiral, 119, 129
  structure, 93–131
  transformation, 210, 229, 234
  Zoo project, 3, 76–77, 79
Galaxy–galaxy encounters, 212
Gas accretion, 202
Gas halo, 213
Gas removal, 213, 216, 220, 224, 225, 228, 235, 250, 257
Gas shadowing, 254
Global Astrometric Interferometer for Astrophysics (GAIA), 426, 446
Globular cluster luminosity function (GCLF), 433
Grand design spirals, 6, 28
Gravitational interactions, 199, 211–213, 215, 225, 228, 229, 255, 256
Gravitational lensing, 269, 270, 281, 297–298
Group environment, 210, 211, 219, 255, 257
Groups of galaxies, 210, 256

H

Halo models, 406–413, 418, 459–460
Halo occupation, 407–410, 418, 420
Harassment, 212, 213, 228–230, 255–257, 482, 497
Hickson compact groups (HCGs), 220, 223, 234
HI deficiency, 219–221, 223, 224, 235
HI-deficient Hickson compact groups, 234
Higher-order correlation function, 414, 417
H_2O (water) masers, 426, 439
Horizontal branch, 428, 430–432, 439
Hubble constant, 436, 439, 445–446
Hubble, E., 424, 433, 434, 436, 439, 441–446
Hubble–Sandage–de Vaucouleurs classification, 8, 14, 27, 69, 71, 81
Hydrodynamical simulations, 465, 467, 471, 474, 484, 490–492
Hydrostatic equilibrium, 289, 296–297

I

IC 3418, 247, 248, 255
IC 3428, 247
Inner rings, 6, 8, 9, 14, 19, 22, 24, 25, 33–35, 39, 53–56, 58
Interaction diagnostics, 231–253
Intergalactic medium (IGM), 568–605
Interstellar medium (ISM), 184–202, 213–215, 223, 230, 231, 237, 247, 251
Intracluster gas, 213, 256, 257
Ionized hydrogen, 184, 192
Irregular galaxies, 14, 20, 67, 68, 83
ISM. *See* Interstellar mediums (ISM)
Isolated galaxies, 31, 77–79

J

Jets, 314–317, 327, 330–335, 342, 366–380, 504, 505, 513, 515, 539–543, 545–555, 557–559

K

Kelvin–Helmholtz instabilities, 214–217, 238
Key project, 443, 445, 446

L

Large magellanic clouds, 426, 436–445
Large-scale magnetic field, 231, 232
Late-type galaxies, 8–10, 37, 55, 57, 66, 80
Leavitt law, 427, 434–436, 440, 442, 446
Lens galaxies, 6, 60
Lensing, 612, 619, 643, 664–668, 670–675, 679–681
Lenticular, 210, 211, 226, 228–231, 234, 255–257
  galaxy, 226, 228, 231
LLS. *See* Lyman limit systems (LLS)
LOFAR, 257
Low-Ionization Nuclear Emission Regions (LINERs), 310–312, 340
Low surface brightness galaxies, 30, 69–70, 82
Luminosity classes, 8, 30, 31, 69
Luminosity evolution, 469, 474, 477, 482, 493
Luminosity function, 277, 281–283
Ly α forest, 575, 576, 578–584, 592, 599, 600, 602, 603
Lyman limit systems (LLS), 575, 584

M

M 32, 230, 231
M 49, 238, 251
M 86, 238, 245, 247
M 87, 221, 236, 238, 242, 245
Mach number, 237, 238
Magellanic barred galaxies, 17, 82
Magellanic clouds, 426–427
Magnetic field, 215, 219, 224, 251
Maser galaxies, 439–441
Megamaser cosmology project, 441
Merger morphologies, 68
Metallicity, 222–224, 237, 586, 588, 589, 592, 594, 595, 597, 600, 604
Metal-line systems, 575, 589–599
Mira variables, 431–432
Missing baryons problem, 601–605
Missing metals problem, 599–601
Molecular hydrogen (H_2), 183, 194, 195, 221–223, 241
Morphological segregation, 211, 228
Morphological transformation, 213, 255–257
Morphologies of galaxies, 283
Morphology-density relation, 211, 226
Multiphase ISM, 219–224
Multiwavelength imaging observations, 231, 235
Multiwavelength photometry, 250

N

NASA/IPAC Extragalactic Database (NED), 424, 426
N-body simulations, 456, 457, 459, 460, 462, 463, 490, 493, 495
Nearby galaxy groups, 232–235
NED-D, 424
Neutral hydrogen (H I), 184–193
  distribution, 184–186, 188–192, 196, 199, 201
  kinematics, 184, 199
NGC 1404, 237, 238

NGC 4254, 242, 245, 246, 255
NGC 4258, 436–441, 444
NGC 4330, 232, 248–250, 252–254
NGC 4388, 247, 249–251, 254
NGC 4402, 224, 232, 251, 252, 254
NGC 4424, 242
NGC 4435, 245
NGC 4438, 212, 224, 238, 242, 245, 247, 249–251, 254, 255
NGC 4472, M 49, 238
NGC 4490/4485, 235
NGC 4501, 248, 250
NGC 4522, 213, 214, 217, 231–233, 245, 249–252, 254
NGC 4548, 255
NGC 4552, 237, 238
NGC 4569, 249, 250, 255
NGC 4579, 255
NGC 4654, 242, 244, 245, 255
NGC 4848, 239, 245, 254
NGC 7319, 232
NGC 7318b, 234
Norma cluster (Abell 3627), 235
Nuclear bars, 6, 24–27, 29, 37
Nuclear rings, 6, 18, 24–27, 29, 34, 37, 39, 49

O

One-sided gas tail, 231, 253
Orbital segments, 250
Outer rings, 6, 21–24, 26, 33, 34, 39, 49–51
O VI absorbers, 592–595, 599, 604

P

Parallaxes, 425, 430, 431, 435, 446
Photoionization, 346, 350–356, 361, 364–366
PIXIE. *See* Primordial inflation explorer (PIXIE)
Planck, 614, 615, 627, 629, 630, 634, 635, 643, 644, 654, 655, 660, 665–670, 672, 673, 675, 676
Planetary nebula luminosity function (PNLF), 432–435
Polarization, 611, 612, 614, 616–620, 627–629, 640–649, 651, 653, 663–680
Polarized radio continuum, 219, 231–233, 239, 248–250, 253, 254, 257
  ridge, 253
Polar ring galaxies, 40, 41, 48
Population III (Pop III) stars, 571–573
Post-startburst galaxies, 210, 226
Preprocessing, 212, 226, 228
Pressure jump, 237
Primordial inflation explorer (PIXIE), 612, 620, 625, 627, 629, 640–643
Pseudobulges, 37, 38, 60, 78, 81
Pseudorings, 6, 19, 21–24, 35, 39, 49–51, 56, 58, 81

Q

Quasars, 310, 312–315, 321, 322, 327, 329, 333, 335, 343, 359, 361–366, 370, 380, 507, 509, 510, 512, 521, 527, 529, 537, 541, 552, 555, 557–561, 573, 574, 579
Quenching of star formation, 210, 228, 230, 257

R

Radio continuum emission, 219, 224, 231–233, 239, 248–250, 254
Radio continuum tail, 239, 242
Radio-deficient region, 231, 232, 239, 248, 253
Radio emission, 307–309, 312–317, 371, 372, 374, 378
Radio emission, polarized
  asymmetric ridges of polarized, 239
Radio/FIR distributions, 231
Radio galaxies, 269, 271, 293–295, 515, 542, 544–547, 551, 552
Radio luminosity functions, 224
Ram pressure, 483–486
  stripping, 201, 213, 215–218, 220, 221, 224, 226, 228–231, 234, 235, 238, 239, 242, 244, 245, 247–258
  wind, 213, 216, 217, 231, 251, 253, 254, 257
Ram pressure, backfall
    backfall of stripped, 217
RB 199, 245, 247, 248, 255
Recombination, 568, 569, 572, 573
Red clump stars, 430–432
Redshift space distortions, 395, 397, 398, 403–405, 413, 418
Reionization, 568, 570–573, 584, 585
Relativistic electron, 224, 231
Resonance ring galaxies, 22, 41, 42, 58, 82
Restoring force, 213
Reynolds number, effective, 213
Rotational parallax method, 426
RR lyrae stars, 427–430

S

SBF method. *See* Surface brightness fluctuation (SBF) method
Scaling relations, 506, 507, 531, 535–540, 555, 559
Schmidt law, 144, 166–175, 178, 225
SEAM. *See* Spectral-fitting expanding atmosphere method (SEAM)
Secondary bars, 6, 24, 37
Semi-analytic models, 492–493, 496, 497

Sérsic model, 93, 98–102, 104, 107, 110, 112, 114, 116, 118, 119, 128, 130, 131
Seyfert galaxies, 307, 321, 356, 357, 359, 363, 366
SFH. *See* Star formation history (SFH)
SFR. *See* Star formation rate (SFR)
S0 galaxies, 8, 11, 12, 14–20, 30, 32, 33, 40, 42, 60, 71, 73, 82
Shell/ripple galaxies, 41, 46–48, 81
SKA. *See* Square kilometer array (SKA)
SMBH. *See* Supermassive black holes (SMBH)
Smoothed particles hydrodynamics (SPH), 215–217, 254
Specific star formation rate (SSFR), 144, 175–177
Spectral-fitting expanding atmosphere method (SEAM), 434
Spectrum distortion, 620, 621, 623, 627, 635, 639, 643
SPH. *See* Smoothed particles hydrodynamics (SPH)
Spiral arm(s), 144, 149, 151, 168, 169, 174
  classes, 28, 29
  multiplicity, 79
Spiral galaxies, 6, 14–19, 29–31, 33, 36–38, 42–44, 47–49, 55, 59, 60, 64, 65, 75, 77, 83
Spirals, 5–9, 14–20, 22–24, 27–33, 36–39, 41–44, 47–49, 51, 53, 55, 56, 58–60, 62, 64–66, 68–72, 74–79, 82, 83
Spitzer space telescope, 446
Square kilometer array (SKA), 257
SSFR. *See* Specific star formation rate (SSFR)
Standard candles, 425, 445
Standard model, 611, 612, 616–620, 623, 648, 664–665, 667
Standard rulers, 425
Starburst, 152–153, 155, 156, 159, 161, 164–166, 172, 174–175, 178
Star formation, 143–179, 184–189, 191, 193–199, 202, 209, 213, 215, 222–226, 228–231, 247, 250, 253, 255–257, 454, 463, 465, 467–471, 473–475, 477, 478, 484, 486, 490, 492, 495
  threshold, 194, 197, 198
Star formation history (SFH), 144, 160, 165, 166, 176–179
Star formation rate (SFR), 143–145, 147–149, 151–161, 167, 168, 176, 177
Star-forming galaxy morphologies, 33, 34, 36, 39, 42–44, 53, 56, 62, 64, 66, 71, 74, 77
Starvation, 213
Stellar mass morphology, 59–65
Stephan's quintet, 234
Sticky particles, 215, 217, 254
Strangulation, 213, 486
Stripping efficiency, 216, 217
Sunyaev–Zel'dovich (SZ), 614, 621, 624, 626, 635, 657, 667, 668, 670, 675
Superluminal motion, 316, 327–333
Supermassive black holes (SMBH), 325, 326, 369, 379–381, 505, 507, 511, 512, 516, 517, 520, 523, 528, 530–540, 542, 559, 560, 563
Surface brightness fluctuation (SBF) method, 433, 441–442, 446
Synchrotron radiation, 373

T

Thermal bremsstrahlung, 285–288
3D velocity vector, 237, 250
Threshold, 146, 147, 165, 168, 169, 175–176
Tidal interactions, 211, 212, 217, 219, 220, 223, 225, 226, 228, 231, 232, 235, 239, 242, 245, 251, 255–257
Tidal shocking, 213
Tidal stripping, 230, 231
Tidal tails, 44–46, 48, 68
Time sequence, 247, 249
Tip of the red giant branch (TRGB) method, 431, 437–442, 445, 446
Toomre parameter, 146, 149
Trignometric parallaxes, 425, 435, 446
Truncated star formation, 250
Tully–Fisher relation, 442, 443
Turbulent viscous stripping, 214–216, 238
Two-point correlation function, 388, 391–393, 395, 397–400, 408, 413, 414, 417–419
Type Ia supernova, 426, 442–446
Type II supernovae, 434

U

UGC 3789, 441
Ultracompact dwarf galaxies (UCDs), 230
Ultra-luminous infrared galaxies (ULIRG), 47–48
Unified AGN models, 325–335
UV disks, 197

V

Variability, 322–327, 331–334, 348, 352, 361, 363, 364, 376–379
Virgo, 210, 212–214, 220, 221, 223–228, 230, 232, 233, 235–238, 242–257
  Abell 1367, 235, 256
  cluster, 210, 212–214, 220, 224–228, 230, 232, 235, 236, 238, 242, 244–247, 249–256
  dwarf galaxies, 230
Viscous stripping, 213–216, 238
VIVA, 242
Void, 389–391, 402, 405, 413–419
Void probability function (VPF), 417–418

W

Warm-hot intergalactic medium (WHIM), 575–578, 599–605
Warped disks, 48–49, 68
Warps, 190, 199
Wilkinson microwave anisotropy probe (WMAP), 569, 573, 592, 601, 612, 618, 619, 622, 627, 629, 634, 635, 646–656, 658–660, 662–665, 670

X

X-ray bright groups, 220, 226, 235
X-ray emission, 268, 285–295, 309, 354, 372, 378, 379
X-ray tails, 241