Jean Louis Guénet · Fernando Benavides
Jean-Jacques Panthier · Xavier Montagutelli

Genetics of the Mouse

Springer

Genetics of the Mouse

Jean Louis Guénet · Fernando Benavides
Jean-Jacques Panthier · Xavier Montagutelli

Genetics of the Mouse

Springer

Jean Louis Guénet
Institut Pasteur (Emeritus)
Paris
France

Fernando Benavides
Division of Basic Science Research
Department of Molecular Carcinogenesis
The University of Texas MD Anderson
Cancer Center
Smithville, TX
USA

Jean-Jacques Panthier
Mouse Functional Genetics Unit
Institut Pasteur
Paris
France

and

Ecole Nationale Vétérinaire d'Alfort
Maisons-Alfort
France

Xavier Montagutelli
Mouse Functional Genetics Unit
Institut Pasteur
Paris
France

ISBN 978-3-662-44286-9 ISBN 978-3-662-44287-6 (eBook)
DOI 10.1007/978-3-662-44287-6

Library of Congress Control Number: 2014945779

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

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

This book is dedicated to
the memory of
Professor François Jacob
(June 1920–April 2013)

Foreword

The science of experimental biology rests on the analysis of causative factors, followed by synthesis. Commonly, the analytic step involves determining the consequences of a known perturbation. Classical experimental biology rested on perturbations of the environment, or on surgical operations such as transplantation. When the science of genetics reached molecular resolution in the twentieth century, mutational perturbation became prominent. In organisms for which sophisticated genetic methods have been developed, it is feasible, either through positional cloning of the mutated gene or through directed mutagenesis, to make connections between changes in phenotype and specific molecular changes. The laboratory mouse is the first experimental mammalian species allowing these sophisticated methods. Thus, *The Genetics of the Mouse* by Guénet, Benavides, Montagutelli, and Panthier is more than a genetics textbook. It is also a talisman, containing instructions by which the experimental mammalian biologist can analyze a process of interest at molecular resolution. It is a twenty-first-century version of the twelfth-century tome on the crafts of the medieval guilds authored by Theophilus: *On Divers Arts.*

The chapters delve deeply into the biology of the mouse. They range from detailed presentations of the natural history of the species, its handling in the laboratory, and its classical genetics, to contemporary issues including the epigenetics of parental imprinting and X-chromosome inactivation. Further, they provide a detailed discussion of the strategies for creating and cloning constitutive and conditional mutant alleles. Finally, they present a platform from which the analysis of complex quantitative traits is currently addressed. When the in-depth details of a subject exceed reasonable limits in length, the authors provide footnotes to more extensive treatments. As experienced geneticists, the authors appreciate the importance of phenotyping, not letting it get lost in the details of analyzing and manipulating the genotype. At the core of their presentation is the importance of inbred strains and isogenicity for the identification of single causative factors.

The ultimate goal of many mammalian experimental biologists is to develop an understanding of issues in human biology. The authors recognize the circumstances in which a particular mouse model fails to present the phenotype expected

from the cognate condition in the human, and they outline ways in which mice can be made chimeric for human tissues. Because any one model gives at best only a first approximation to the human case, a diverse set of models may provide further approximations. The methods presented can lead to the development of a homologous series of mouse models in any of their distinct inbred backgrounds, or in their genetically homogeneous F1 hybrids, or in other mammalian genera that can be inbred.

Seen broadly, *The Genetics of the Mouse* connects the past, present, and future in the experimental biology of mammals.

William F. Dove
Streisinger Professor of Experimental Biology, Emeritus
University of Wisconsin
Madison

Preface

This book is intended for several different categories of potential readers. First, are students who have completed their university studies in biology or medical sciences and wish to undertake a PhD project making use of mice but who have no experience with this model organism. Reading this book will enable them to acquire, rapidly and in a relatively condensed form, a background that will be helpful for the critical reading of primary scientific publications and for the optimal design of their projects. Genetics instructors will also find useful examples to illustrate undergraduate biology courses. Molecular and developmental biologists whose research program is focused on a gene or gene family will also be interested and will realize that the mouse is an exceptional model with which they may be able to develop studies impossible or difficult to achieve with any other mammalian species. For example, they may be able to produce a variety of point mutations in the same genetic background or exactly the same point mutation in a variety of different backgrounds, allowing exploration of the function of this gene and its interplay within gene networks. This book will also be helpful to physicians and pediatricians by allowing them to choose or design the best possible model for their research related to a specific human pathology. This would be true not only for the diseases resulting from point mutations in orthologous genes but also, and more interestingly, for those mutations whose phenotypic expression is influenced by the environment or the genetic background of the animal. Finally, laboratory animal veterinarians and technicians, who are in charge of the breeding and preservation of mouse models, will find useful explanations about their increasing complexity.

This book covers all aspects of mouse genetics. The first four chapters describe the origin of laboratory mice, the reproductive biology, the cytogenetics, and the mapping of genes. The establishment of highly detailed genetic maps was a major and fundamental contribution to mouse geneticists during the twentieth century that ultimately led to the complete sequencing of the genome. This topic has been presented in a relatively condensed form in this book, because we have considered that the excellent book published in 1995 by Lee M. Silver, which is freely available on the site “Mouse Genome Informatics”, is still a major reference in

this matter. On the contrary, the transcriptome and the parental imprinting of the genome are topics that have been the subject of intensive research over the last 10 years. For this reason they are presented in more detail along with the techniques for the production of mutations, which is one of the most attractive features of the mouse. Finally, quantitative genetics, a branch of genetics that is in expansion, is presented in a didactic manner.

This book greatly benefited from the contributions of some of our colleagues whom we would like to cordially thank. François Bonhomme, an old friend with whom we have collaborated many times in the past, reviewed and commented on Chap. 1. Marie-Geneviève Mattei read and amended Chaps. 3 and 6 and allowed us to share her extensive knowledge of cytogenetics. Yann Herault also made interesting suggestions about Chap. 3 and provided us with a schematic figure representing the best models of Down syndrome. Benoît Robert accepted the difficult task of writing an original synthesis concerning the regulation of gene expression (Chap. 5). Edith Heard, Luisa Dandolo, and Deborah Bourc'his abundantly corrected and commented on Chap. 6 dealing with X-inactivation and parental genetic imprinting. Michel Cohen-Tannoudji corrected and completed our initial versions of Chap. 8, and Tomoji Mashimo read the section of the same chapter dealing with the production of targeted alterations using engineered nucleases and provided a summary picture. Finally, Robert P. Erickson kindly read the whole of our manuscript, making many insightful comments. The authors also wish to thank Drs. Hesed M. Padilla-Nash and Thomas Ried from the Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda for providing a picture of a mouse spectral karyotyping, Dr. Dianne Creasy, Huntingdon Life Sciences, East Millstone, for providing a picture of the seminiferous epithelium with identification of the different cell types, and Ms Annie Orth for providing a picture of a sample of her unique collection of wild mice. Finally, the authors are greatly indebted to their colleague Dominique Simon, who helped in the preparation of many illustrations and to Mrs. Sarah Adai, MD Anderson Cancer Center, who undertook to “translate” their awkward English into a more readable form.

Writing this book has kept us busy for nearly two years, but it was really an enthralling experience. Whatever the chapter, we realized that the Genetics of the Mouse has changed considerably over the last 20 years and, with an increasing number of transnational collaborative projects, we can expect even more dramatic changes in the years to come.

Contents

Chapter 1
Origins of the Laboratory Mouse

1.1 Introduction

Because they are often closely associated with humans and sometimes "share" food with them, zoologists consider the mouse as a commensal species (from the Latin *cum mensa*, which means eating at the same table). For the same reason, mice are often referred to as "house" mice in the English literature, as opposed to wild or feral mice even though, in fact, they are the same species. Because they have been described as "invasive," "prolific," "troublesome," and "devastating," farmers consider mice to be pests. Physicians and epidemiologists don't like mice (and rodents more generally) because they are natural reservoirs of many pathogens, some of them deadly. For some other people, on the contrary, mice are cute pets, easy to breed, cheap to buy, and with beautiful coat colors (Fig. 1.1).

Researchers, geneticists in particular, have the greatest respect for mice and have even graded the species to the rank of *domestic* species, which they breed in large numbers to fulfill their needs of experimental models. With so many divergent opinions, the time has come to address a few basic questions about mice: what are they actually? Where do they come from? Why and how have they become such popular models for research in genetics over the last century?

1.1.1 Phylogenetic Relationships of Laboratory Mice with Other Mammals

Figure 1.2 represents a phylogenetic tree of a sample of 28 vertebrate species, including most domestic species (dog, cow, chicken, etc.) as well as species that are commonly used in scientific research (mouse, frog, zebrafish, etc.). This evolutionary tree is a useful tool for molecular biologists wishing to make comparisons at the genomic level (sequence comparisons, origin of gene families, etc.). We will frequently refer to it in this book.

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_1

Fig. 1.1 *Some* **mouse** *mutations with effect on the coat color*. Mice represented in this figure are homozygous or heterozygous for mutations affecting coat color. Many phenotypes of this kind have been collected over the years by fanciers and geneticists and are still for sale in many traditional pet shops. Associated or not in the same individual, they have produced a large variety of beautiful specimens that have captured the enthusiasm of many children. Even today, rare specimens are regularly exchanged between members of many pet clubs. Coat color mutations, behavioral mutations, and mutations affecting the fur or the skeleton were also used by mouse geneticists in the early days as genetic markers for the detection of genetic linkage because they had little or no effect on viability and fertility. The first linkage discovered in the mouse (linkage group I—now chromosome 7) was between the coat color mutations *pink-eyed dilution* (*p*, now *Oca2*^{p}) and *chinchilla* (*c*^{ch}, now *Tyr*$^{c\text{-}ch}$)

Laboratory mice (together with rats and Guinea pigs) belong to the order *Rodentia,* which is the largest group of mammals on earth, comprising around 40 % of mammalian species (Fig. 1.3). Rabbits are not rodents *sensu stricto* but lagomorphs[1]. However, due to their evolutionary proximity they are often merged with the rodent family and together with them are referred to as the superclass *Glires*.

About two-thirds of rodent species belong to the superfamily *Muroidea*, a superfamily that is itself composed of six families including the *Muridae* family, which includes the "house" mouse *Mus musculus*[2] and two species of rats, *Rattus norvegicus* and *Rattus rattus*. This is a very large family of mammals with at least 1,300 species. Laboratory mice belong to the genus *Mus* that itself contains four subgenera: *Mus*, *Coelomys*, *Pyromys* and *Nannomys*, and at least 40 different

[1] There are two families in the order *Lagomorpha*: the *Leporidae* (hares and rabbits) and the *Ochotonidae* (pikas).

[2] Many rodent species carry the name "mouse", meaning a mouse-like small furry creature, hence the importance of the binominal nomenclature.

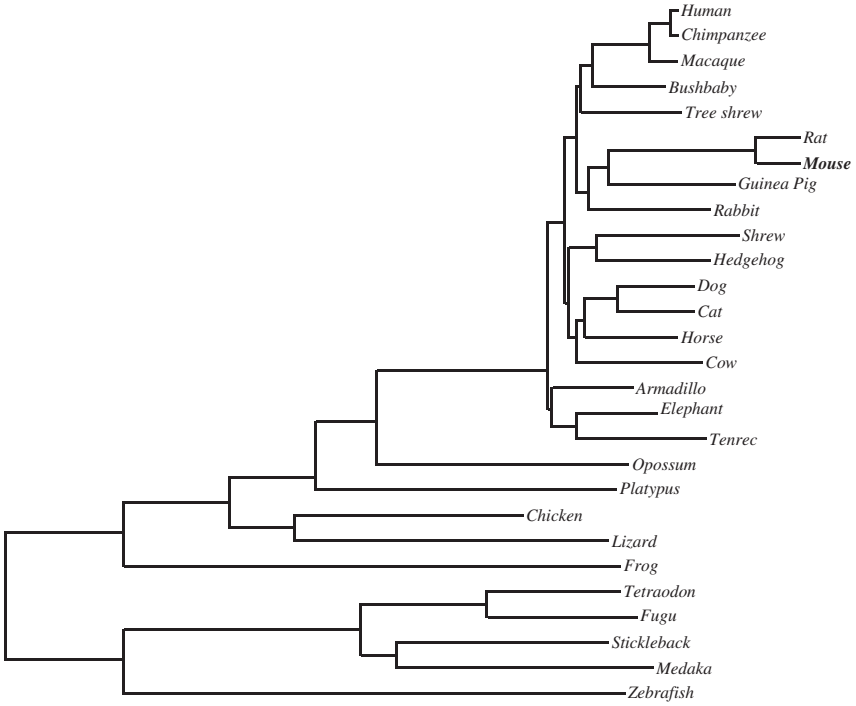


Fig. 1.2 *Phylogenetic tree representing the relationships between 28 vertebrate species*. Branch lengths are proportional to the number of base-pair substitutions at a number of specific sites. The estimated time of divergence between humans and mice is approximately 75–80 Myr ago (*redrawn and modified from* Blanga-Kanfi et al. 2009)

species. Figure 1.4 summarizes the phylogenetic relationships within the genus *Mus*. Inside each subgenus the different species are, with few exceptions, extremely similar in size and morphology.

Because of these similarities, the phylogenetic relationships between the different species have been difficult to establish, especially when morphological characteristics (tail length, body shape, coat color, habitat, etc.) were the only criteria taken into account for the establishment of the systematics (Fig. 1.5). Nowadays, with possible reference to the complete genomic sequence of many genes as well as to the sequence of mitochondrial DNA, the situation has been much clarified.

The geographic distribution of the genus *Mus* encompasses all of Eurasia and Africa. The presence elsewhere of a single of its species, the house mouse, particularly in Australia and the Americas, results from human-mediated introductions during recent centuries (Jones et al. 2013).[3] Hence, the house mouse

[3] The house mouse, which is now endemic in Australia and the Americas, was involuntarily transported from Europe or from Asia by maritime traffic. Many genetic markers (endogenous copies of retroviruses inserted as proviral DNA, in particular) confirm the origin of these "stowaways".

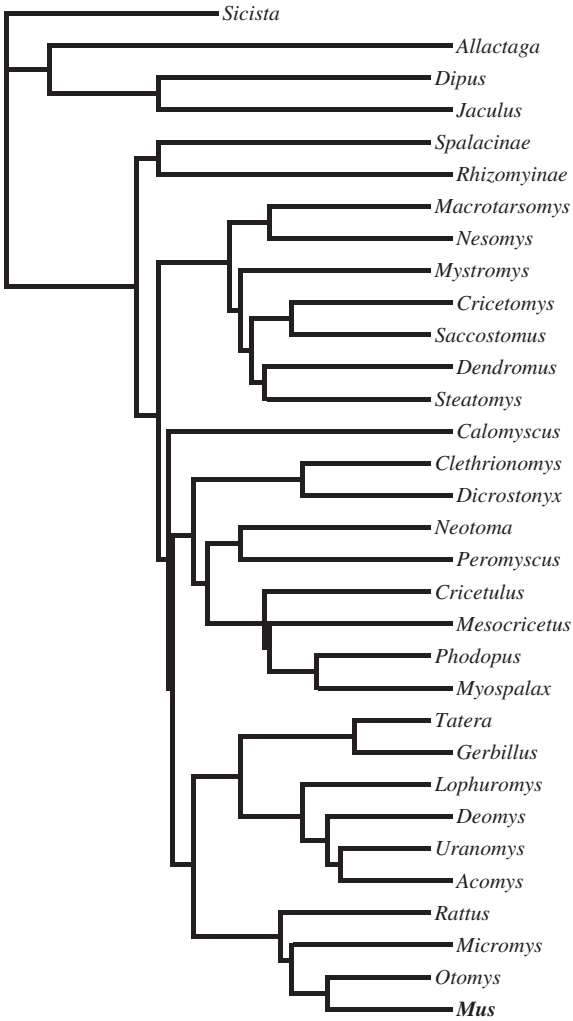


Fig. 1.3 *Phylogenetic relationships between 32 species of rodents representing 14 subfamilies of the Muridae family*. The estimated time of divergence between the mouse and rat species is approximately 12/15 Myr ago (redrawn from Michaux et al. 2001)

is currently widely spread over the five continents, with the highest diversity in Asia (with 3 subgenera and ~20 species), where this genus likely originated. Based on recent observations, and if we consider that the habitat of some (still unknown) species might be very limited, and possibly embedded in the wider habitat of other species, it is likely that the number of species in the subgenus *Mus* will increase further (Bonhomme et al. 2004).

The evolutionary divergence between humans (*Homo sapiens*) and mice of the *Mus* genus probably occurred 70–75 million years (Myr) ago (Fig. 1.2) while the divergence between humans and the other domesticated species (e.g., dog, cat, horse and cow) is slightly greater (80–85 Myr) (Murphy et al. 2001). The divergence between the *Mus* and *Rattus* genera probably occurred around 10–12 Myr ago. Finally,

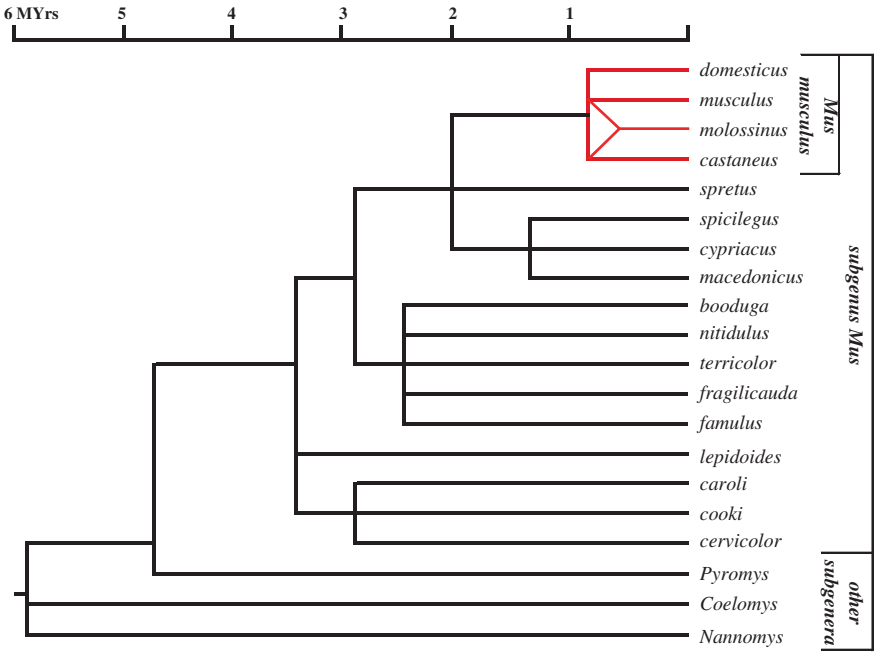


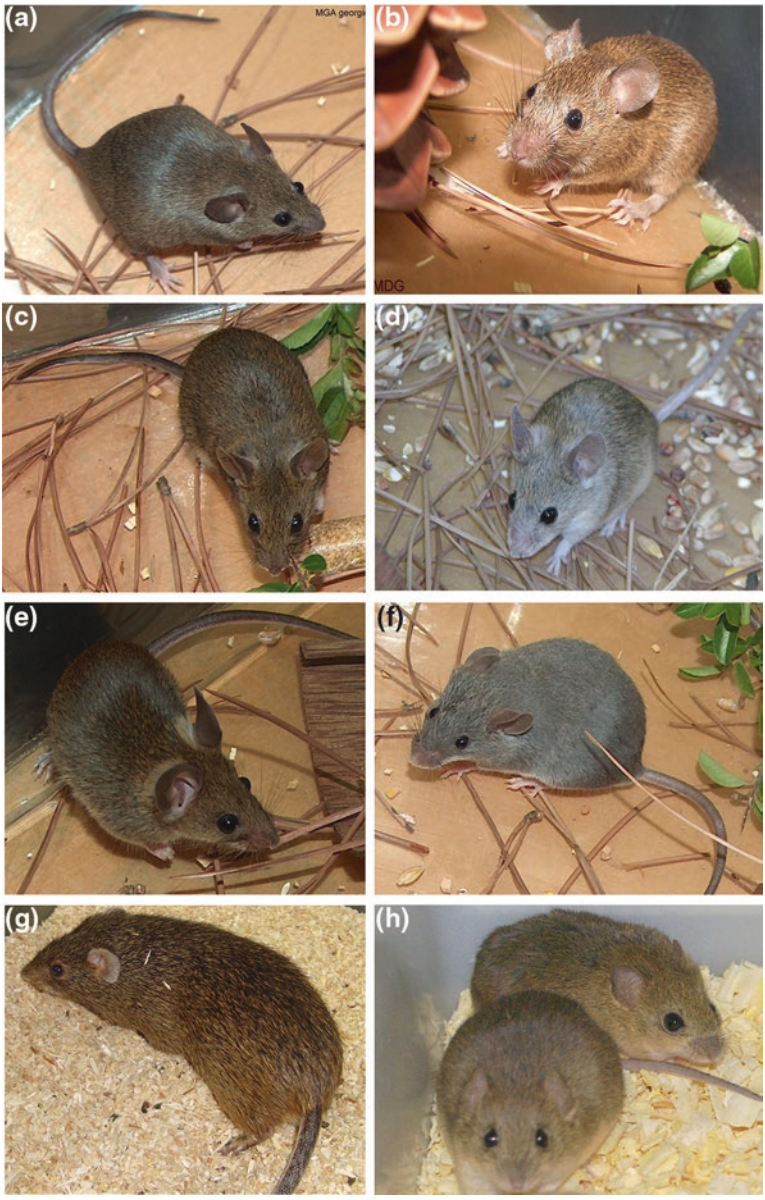
Fig. 1.4 *Consensus phylogenetic tree of the genus Mus issued from a compilation of all existing studies*. The estimated time of divergence of the different *Mus* species is indicated at the *top* of the diagram

the individualization of the subgenus *Mus* sensu stricto occurred around 6 Myr ago with the split from three other subgenera (Boursot et al. 1993; Musser and Carleton 1993; Chevret and Dobigny 2005; Chevret et al. 2005; Suzuki and Aplin 2012).

1.1.2 How the House Mouse Became a Domestic Species

The beginning of human/mouse commensalism is very ancient and probably dates to some 12,000 years ago. Mouse remnants and paintings have been discovered in the very large Neolithic settlement of Çatal Hüyük (southern Anatolia, Turkey), which existed from approximately 7,500 to 5,700 B.C., suggesting that, during this period, mice were considered (worshipped?) as holy creatures or living symbols. Finally, numerous historical records (Keeler 1931; Morse 1978; Berry 1987; Sage et al. 1993; Moriwaki et al. 1994) indicate that 3,000 years ago mice were pet animals in Europe, Japan, and China.[4]

[4] In Japanese traditional writing there is only one Kanji to define both rats and mice: *nezumi*. This is a possible source of confusion.


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

◀ **Fig. 1.5** *Some specimens of the order Rodentia*. This panel represents eight specimens of the order *Rodentia*. In spite of great similarities in size and body shape, some of these "mice" are only weakly related species. *Mus m. castaneus* (**b**) and *Mus spretus* (**c**) can produce viable and fertile hybrids with mice of the *Mus m. domesticus* species (**a**) or with laboratory strains. Interspecific hybrids resulting from crosses between *Mus spretus* males and laboratory females are fertile but only in the female sex (Haldane's rule), and this sort of cross has been used extensively for the development of the mouse genetic map. The reciprocal cross (laboratory males × *Mus spretus* females) is much less fertile and produces hybrids only in special conditions. The possibility of obtaining hybrids between *Mus cypriaticus* (**d**) and laboratory strains has not yet been tested. Hybrids generated by the artificial fertilization of laboratory females with sperm of *Mus caroli* (**e**) complete fetal development, and a low percentage of them survive to maturity but are stunted and do not reproduce. Embryonic cells of *Mus caroli* can participate in the formation and development of a chimeric fetus when associated with cells of a laboratory inbred strain. Hybrids between *Coelomys pahari* (**f**) and laboratory strains have never been produced and would presumably not be viable. Rodents of the *Arvicanthis ansorgei* species, also known as the Sudanian grass rat (**g**), are endemic in West African countries and do not produce hybrids with mice of the genus *Mus*. Rodents of this species, unlike the other rodents presented here, have essentially diurnal activity. Finally, *Calomys callosus*, the large vesper mouse, is a South American rodent of the family *Cricetidae*. Despite their similarities to the other mice represented in the picture, which all are of the family *Muridea*, these rodents are phylogenetically closer to hamsters (*Cricetulus griseus*) and deer mice (*Peromyscus maniculatus*) than to mice of the genus *Mus*. Several of these species and subspecies have been established as laboratory colonies. One of the most diverse collections of wild-derived strains can be found at the Université de Montpellier, Place Eugène Bataillon, France, c/o Dr. François Bonhomme. Six pictures in this panel (**a**–**f**) are from the wild rodent repository of Dr. François Bonhomme. The picture of *Arvicanthis ansorgei* (**g**) is from Dr. Sophie Reibel-Foisset, (Chronobiotron, Strasbourg, France). The picture of *Callomys callosus* (**h**) is from Dr. Adriano Abbud (Instituto Adolfo Lutz, São Paulo, Brazil)

All this evidence indicates that mice and humans have been in contact for a very long time. It was then logical that these small mammals, as well as the rat, were used by early scientists for performing their experiments, and if this choice appears nowadays to be more opportunistic rather than based on scientific considerations, it nevertheless appeared to be an excellent one in the context of modern biomedical research.

Mice are easy to breed. As they are rodents, they eat a rather large quantity of food but do not have very specific or expensive nutritional requirements. When kept in laboratory facilities with stable environmental conditions (light and temperature), they do not hibernate (meaning that they have a decreased physiological activity) and breed all year round, with a short generation time. They deliver relatively large progenies and tolerate inbreeding rather well compared to other mammalian species. For all these reasons, but also because some ancestral specimens were tame and easy to handle, mice have been used in biomedical research since the beginning of the sixteenth century, when biology gradually shifted from a descriptive to an experimental science. Herbert C. Morse (1978, 1981) reported that William Harvey (1578–1657) used mice for his fundamental studies on reproduction and blood circulation while according to Richard J. Berry (1981), the earliest record of the use of mice in scientific research seems to have been in England, in 1664, when Robert Hooke (1635–1703) used mice to study the

biological consequences of an increase in air pressure. Much later, Joseph Priestley (1733–1804) and his intellectual successor, Antoine Lavoisier (1743–1794), both used mice repeatedly in their experiments on respiration. Subsequently, an ever-increasing number of scientists used mice in their experiments, at least as long as the small size of the rodents was compatible with the experimental project. This is how the house mouse progressively and logically became "the laboratory mouse".

1.1.3 How the House Mouse Became a Model for Geneticists

1.1.3.1 From Louis-Théodore Coladon to Gregor Mendel and Lucien Cuénot

Over the past two centuries, fanciers in Europe, in Japan, and in the United States were regularly breeding and exchanging pet mice with a wide variety of coat color or amusing behavior (e.g., the famous "dancing mice" homozygous for the *Cdh23*^{v} or *waltzer* mutant allele). All these mice were crossed to produce new eye-catching phenotypes to supply the market, and it progressively became obvious that many of these traits were inherited. The mouse then appeared to be an excellent model for studying inheritance!

According to Hans Grüneberg (1952) and Jean Rostand (1957), some among these fanciers played an important role. Louis-Théodore Coladon (sometimes spelled Colladon; 1792–1862), a pharmacist established in Geneva, was probably the first to report the results from breeding experiments achieved between 1825 and 1829, which were in perfect agreement with what we now call the *Mendelian ratios*. This, however, was 36 years before the publication of Mendel's own results on peas. The experimental results of Coladon have never been published but were merely quoted by Jean-Baptiste Dumas, the famous chemist, in the *Dictionnaire classique d'Histoire naturelle* (tome 7, p. 202) and by William F. Edwards, a physiologist and ethnologist, in a book published in 1829.

As revealed by Hugo Iltis (1932) and commented on by Kenneth Paigen in his notes on the history of mouse genetics (2003a, b), we know that Gregor Mendel also bred mice (*grey* and *white* mice) in his monastic room. These mice were bred as pets, but it is likely that Mendel (who later proved to be a sagacious observer!) had his attention focused on the segregation of coat color in the progenies and may have had sufficient experimental data to make observations on the transmission of these characteristics. However, Mendel never commented on these observations and was asked by his hierarchy to stop breeding mice. At the time it was positively indecent and even immoral to perform experiments dealing with animal reproduction (mating) and inheritance, particularly in a monastery. Accordingly, Mendel changed his experimental model to garden peas and published his famous observations in 1866 in a botanical journal, where they had a rather low impact and remained virtually ignored until the beginning of the twentieth century.

Once rediscovered by H. de Vries, C. E. Correns, and E. von Tschermark-Seysenegg, the three of them working independently with plants, it was tempting to test whether the so-called Mendel's laws were also valid for animals.[5] Lucien Cuénot (1902), a professor of biology at the University of Nancy (France), crossed mice segregating for several common coat color markers including the albino (*Tyr*^{c}) and published the results of experiments indicating that this was indeed the case. Cuénot's observations were shortly confirmed and extended to other species, as well as for other genetic traits by W. Bateson, E. R. Saunders, A. Garrod, W. E. Castle and C. C. Little (Paigen 2003a, b). Cuénot was also able to interpret correctly the unusual pattern of transmission (1/3–2/3 instead of the classical 1/4–3/4) of the yellow dominant allele (*A*^{y}) at the *Agouti* locus (chromosome 2 [Chr 2]), suggesting that *A*^{y}/*A*^{y} embryos died in utero at an early stage of development, and accordingly that yellow mice (*A*^{y}/*A)* could not "breed true" (Cuénot 1905).

1.1.3.2 The Origins of Laboratory Mouse Strains

The majority of albino mouse strains used today in experimental research are derived from ancestral breeders bought in pet shops, which were bred either by the researchers themselves or by amateurs as a source of income. For many years, and even today, many of these albino mice bred for general purpose in laboratories, were collectively designated "Swiss" mice to recall their Helvetian origin (perhaps they were indeed distantly related offspring of Coladon's mice?). These mice were bred with no specific mating protocol, and the only criteria for selecting the breeders, generation after generation, were docility and good health. The breeding colonies were regularly decimated by outbreaks of infectious diseases or sometimes reduced to a few breeding pairs as a consequence of a lack of space (or of funding!). A consequence of this "bottleneck effect" was that the mice became progressively (and insidiously) inbred. However, strict inbreeding was absolutely avoided based on the negative experience of livestock and dog breeders.

Strain DBA/2 (formerly *dba*, then DBA) is the most ancient of all inbred strains. It was started by Clarence C. Little in 1909 (Russell 1978) by intercrossing mice homozygous for the coat color markers non-agouti (*a*), brown (formerly *b*, now *Tyrp1*^{b}) and dilute (formerly *d*, now *Myo5a*^{d}). About 10 years later, Miss Abbie Lathrop of Granby, a retired school teacher from Massachusetts (USA), established strain C57BL/6 by intercrossing the "black" offspring of female 57 (Strong 1978). According to several historical records, Miss Lathrop played an important role in the development of laboratory strains because she was keeping excellent records of the pedigrees of her strains. In collaboration with researchers on the East coast of the United States (in particular, Leo Loeb,

[5] For an interesting historical account, refer to *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics*, by Robin Marantz Henig (2001).

a pathologist working at the University of Pennsylvania), she made interesting observations about the occurrence of specific cancers. Miss Lathrop was well aware of the importance of breeding top-quality mouse strains for the progress of science, and her reputation was so good that the US government made an appeal to her to supply mice (and guinea pigs) to research laboratories (Shimkin 1975; Steensma et al. 2010). Being close to the Bussey Institute at Harvard University, she also collaborated with William Castle (considered the father of mammalian genetics) and Clarence C. Little.

Strains C3H, CBA, and A were created during the same time period by Leonell C. Strong, a cancer geneticist established at Cold Spring Harbor Laboratory (Strong 1978). At this point it is interesting to note that, among the strains established by Leonell C. Strong, strains CBA and C3H stemmed from the offspring of an outcross with wild specimens trapped in a pigeon coop in Cold Spring Harbor. This probably explains how the wild-type allele at the agouti locus (*A*) was reintroduced into laboratory strains.

With a few exceptions, historical records concerning the genealogy of most laboratory inbred strains are well documented, and several interesting reviews on this subject are available (Strong 1978; Festing 1979; Rader 2004; Artzt 2012). A chart describing the genealogy of these strains, including the recently established ones, has been published, and regularly updated information is available from The Jackson Laboratory website (Beck et al. 2000). In addition to the chart published by Beck and co-workers, which was based mostly on historical records, a mouse "family tree" was also published by Petkov and coworkers (2004), which is based on a set of 1,638 informative single nucleotide polymorphism (SNP) markers (see Chaps. 4 and 5), located 1.5 Mb apart and tested in 102 mouse strains. These family trees have been documented further and greatly enriched, and have become invaluable tools for researchers who are willing to make interstrain comparisons because they make it possible to select pairs of strains that are more or less distantly related before comparing specific phenotypic traits (Yang et al. 2007; Szatkiewicz et al. 2008). This is extremely important, for example, for the analysis of quantitative traits (see Chap. 10).

1.1.3.3 Mice Have Been Instrumental in Research in Biology and Genetics

Laboratory mice have been at the origin of many important discoveries in biology. To cite just a few, we could say that our understanding of the genetic determinism underlying the success or failure of tissue transplantations is a direct consequence of experiments performed with inbred mouse strains by Peter Gorer (1948), then by George D. Snell and co-workers (1978). These researchers developed a series of congenic resistant strains that were all genetically identical to the C57BL/10Sn background strain, with the exception of single short chromosomal regions determining graft rejection. These very clever experiments led to

the establishment of the so-called "laws of transplantation" and opened the way to what has become known as *Immunogenetics*. For this discovery "*concerning genetically determined structures on the cell surface that regulate immunological reactions"*, George D. Snell, from the Jackson Laboratory, was awarded the Nobel Prize in Physiology or Medicine in 1980, jointly with Professors Baruj Benacerraf and Jean Dausset.

The hypothesis, proposed by Mary F. Lyon, that one X-chromosome out of two was inactivated in female mammals followed from the observation of variegations in the coat color for some X-linked mouse mutations and was interpreted by using X-autosome translocations (Lyon 1961). Chimeric organisms were produced for the first time by A.K. Tarkowski in Warsaw (1961) and B. Mintz in Philadelphia (1962) by merging in vitro independent mouse embryonic cells.[6] The testicular terato-carcinomas, which are common in strain 129, and the cell lines derived from these tumors and cultivated in vitro, have been a material of choice for investigating the processes at work in tissue differentiation for almost a decade (Stevens and Little 1954; Stevens 1970; Jacob 1983). This work undoubtedly opened the way to the establishment of so-called embryonic stem cells (ES cells) by Evans and Kaufman (1981) and Martin (1981). These ES cells paved the way for the "*discoveries of the principles for introducing specific gene modifications in mice*", for which M.R. Capecchi, M.J. Evans and O. Smithies were awarded the Nobel Prize in Physiology or Medicine in 2007.

The discovery of parental imprinting of some chromosomal regions was a consequence of experiments performed by McGrath and Solter (1984) and Surani and co-workers (1984), who demonstrated that a normal mouse embryo can only develop from the fusion of a male and a female pronucleus, while Cattanach and Kirk (1985) demonstrated that the parental origin of the two elements of a given chromosome pair was not always genetically equivalent.

The first transgenic mammal created by pronuclear injection of cloned DNA was a mouse (Gordon et al. 1980), as was the first mammalian organism genetically engineered in vitro (Kuehn et al. 1987). Only the first cloned mammal was not a mouse, but this type of uniparental procreation has been achieved in the mouse, although the efficiency of the procedure is very low, like in other mammals (Wakayama and Yanagimachi 1999). The first mammal whose genome was completely sequenced was a mouse of the C57BL/6 inbred strain (Waterston et al. 2002).[7] Finally, and to cite just another example among many others, we could say that the discoveries made by Bruce Beutler about innate immunity, for which he was awarded the Nobel Prize in 2011, were made possible by the existence of a large number of mutations induced in the mouse genome by the chemical mutagen Ethyl-Nitroso-Urea.

---

[6] In the 1970s, these chimeric mice were sometimes called *allophenic* to recall their origin.

[7] A draft sequence of the human genome was published 2 years (2000) before the draft sequence of the mouse (2002), but the human sequence still has some gaps while the mouse sequence is 99.5 % complete.

1.1.4 The Community of Mouse Geneticists

As is often the case when independent researchers use the same experimental "material" and the same logistics, a community of mouse geneticists formed over the years at the international level with the mouse as a common denominator. The community had its own journal called *Mouse News Letters* and its own meetings organized at various places in the Northern hemisphere, alternately in Europe, in the USA and sometimes in Japan.

Mouse News Letters, first issued in 1949, was published regularly every semester until 1997 (95 issues). This informal publication, edited by scientists from the Medical Research Council (first at Edinburgh, then at Harwell), was distributed free of charge worldwide for several decades and was the best medium for the dissemination of information among the community. The name *Mouse News Letters* was changed to *Mouse Genome* in 1990, when this publication became a peer-reviewed journal. Finally, in 1998, *Mouse Genome* merged with *Mammalian Genome*—edited and published by Springer Verlag.

Mouse News Letters will forever remain the best place to find information about the history of mouse genetics, and in particular about the history of most traditional inbred strains, the progressive development and refinement of the linkage map, and the discovery and initial description of hundreds of spontaneous mutations. The scientific content of the successive issues of *Mouse News Letters* will never be obsolete. On the contrary, it is the "memory" of the early days of mouse genetics.

1.1.5 The Main Institutions Involved in Mouse Genetics

The Jackson Laboratory (internationally known as the JAX-Lab), which was founded in 1929 by C.C. Little in Bar Harbor (Maine, USA), has played a central role in the promotion of the mouse as a laboratory model and still is the world's largest center for mouse genetics. A second "JAX-Lab" opened recently in Sacramento, California. The Jackson Laboratory is a non-profit organization entirely and exclusively dedicated to basic research on mouse models of human diseases. Its mission is to discover the genetic basis for preventing and treating human disease, and to enable research and education for the global biomedical community. It is, at the same time, a research institution, a meeting place where courses and conferences are organized on various aspects of mouse genetics, and the world's largest genetic repository where a great variety of genotypes (around 6,000) and biological samples of all kinds are stored and distributed to the scientific community either as living animals or, in most instances, in the form of frozen embryos or sperm cells. The Jackson Laboratory is also the home of the Mouse Genome Informatics database (MGI at http://www.informatics.jax.org/), an essential tool for mammalian geneticists.

Several other institutions, like the Oak Ridge National Laboratory in Tennessee (USA) and the MRC centre at Harwell in England, have also played (and still play) a very important role in the development of the mouse as a laboratory model for research in genetics, oncology, and immunology. These two centers were founded after World War II when the British and US governments decided to evaluate the genetic hazards that might be associated with the use of radiation, and, more generally, of nuclear energy. Thousands of mice have been experimentally irradiated in these centers to assess the genetic damage of various types of radiation distributed at various doses. Accordingly, a very large number of mutations and chromosomal rearrangements have been induced, collected, and preserved. Both the mutations and the chromosomal rearrangements have been invaluable tools for the establishment of the mouse genetic map. Many of them are also interesting models of human diseases.

Other research centers must also be mentioned for their contribution to the development of mouse genetics in the second half of the twentieth century: the MRC centre in Edinburgh, Scotland, the Deutsches Forschungszentrum für Gesundheit und Umwelt, at Neuherberg, Germany (now Helmholtz Zentrum München), and the Institute of Genetics at Mishima in Japan.

More recently, the European Union has decided to support the establishment of a wide network of genetic repositories (the so-called European Mouse Mutant Archive or EMMA), with major nodes in Italy (EMMA headquarters is in Monteretondo, near Rome), England (Harwell), France (Orléans-la-Source), Germany (Munich), and Spain (Madrid). Finally, and even more recently, Japanese scientists have created and implemented a large bioresource centre at the RIKEN Institute in Tsukuba, with teaching and research activities focused on mouse embryology and genetics. More information about all these centers is available on their websites.

1.1.6 Books and Other Sources of Information Concerning the Mouse

Readers who are interested in the history of mouse genetics are invited to consult the following books, which are available online at the MGI website.

- *Biology of the Laboratory Mouse* edited by Earl L. Green—Dover Publications 1966
- *Origins of Inbred Mice* edited by Herbert C. Morse III—Academic Press 1978
- *Mouse Genetics—Principles and Applications* by Lee Silver—Oxford Press 1995

Some parts of the book *Biology of the Laboratory Mouse* are obsolete, but many others are still a rich source of information with many references. This book is an excellent textbook for all issues related to linkage and gene mapping.

Four other books are also an important source of information from a historical point of view:

- *Genetics of the Mouse* by Grüneberg—Martinus Nijhoff 1952
- *Inbred Strains In Biomedical Research* by Festing—Macmillan Press 1979
- *Biology of The House Mouse* Edited by R. J. Berry—Academic Press 1982
- *Making Mice: Standardizing Animals for American Biomedical Research, 1900–1955* by Rader—Princeton University Press 2004.

Finally, a number of Websites are commonly used by mouse geneticists. The following list is not intended to be comprehensive and additional URLs will be given in the subsequent chapters:

- Emouseatlas (http://www.emouseatlas.org/emap/home.html) encompasses a 3-D anatomical atlas of mouse embryo development and a database of mouse gene expression.
- Pathbase provides a searchable database of histopathology images derived from experimental manipulation of the mouse genome or experiments conducted on genetically manipulated mice. It is a reference/didactic resource covering all aspects of mouse pathology.
- e!Ensembl (http://www.ensembl.org/Mus_musculus/Info/Index) produces a genome database for the mouse and makes this information freely available online.
- The International knockout consortium (http://www.knockoutmouse.org/) provides information on conditionally trapped and targeted genes in mouse embryonic stem (ES) cells.
- MouseMine (http://www.mousemine.org/mousemine/begin.do) is a powerful system for online access to mouse data from Mouse Genome Informatics.

In the UK, the Sanger Institute Mouse Genetics Project has recently launched the Mouse Resources Portal (http://www.sanger.ac.uk/mouseportal/) with extensive genotyping and phenotyping resources.

1.1.7 The Future of Mouse Genetics

The state of mouse genetics, as expected, changed dramatically at the turn of the millennium for two main reasons. First, because the complete sequence of the genome was established and immediately made available to the community, making possible all sorts of comparisons and predictions at the genome level. In several chapters of this book we will discuss the consequences of this unprecedented achievement and the projects that followed from it. The second reason is that geneticists have now, at their disposition, all the tools and strategies allowing them to make, almost at will, any type of alteration in the genome, from large segmental deletions or additions to single base-pair substitutions. Enthralling projects planned for the years to come have a new dimension, and the “mouse

community" now involves virtually all biologists on the planet using mammals in their research. Many projects of great importance will be undertaken in the future, in particular for understanding the determinism of complex traits. For these projects, no species can seriously compete with the mouse, and this is why we predict a promising future for mouse genetics.

Acknowledgements The authors thank Doctor François Bonhomme, Université de Montpellier, France for his contribution to this chapter as well as for Fig. 1.4.

References

Artzt K (2012) Mammalian developmental genetics in the twentieth century. Genetics 192:1151–1163

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM (2000) Genealogies of mouse inbred strains. Nat Genet 24:23–25

Berry RJ (1981) Biology of the house mouse. Academic Press, London

Berry RJ (1987) The house mouse. Biologist 34:177–186

Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D (2009) Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol 9:71

Bonhomme F, Orth A, Cucchi T, Hadjisterkotis E, Vigne JD, Auffray JC (2004) Découverte d'une nouvelle espèce de souris sur l'île de Chypre. C R Biol 327:501–507

Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F (1993) The evolution of house mice. Annu Rev Ecol Syst 24:119–152

Cattanach BM, Kirk M (1985) Differential activity of maternally and paternally derived chromosome regions in mice. Nature 315:496–498

Chevret P, Dobigny G (2005) Systematics and evolution of the subfamily Gerbillinae (Mammalia, Rodentia, Muridae). Mol Phylogenet Evol 35:674–688

Chevret P, Veyrunes F, Britton-Davidian J (2005) Molecular phylogeny of the genus Mus (Rodentia: Murinae) based on mitochondrial and nuclear data. Biol J Linn Soc 84:417–427

Cuénot L (1902) La loi de Mendel et l'hérédité de la pigmentation chez les souris. Arch Zool exp gén 3e séries 3:xxvii-xxx

Cuénot L (1905) Les races pures et leurs combinaisons chez les souris (4^{éme} note) Arch Zool exp gén 4e séries 3:cxxiii-cxxxii

Evans M, Kaufman M (1981) Establishment in culture of pluripotent cells from mouse embryos. Nature 292:154–156

Festing MF (1979) Inbred strains in biomedical research. The MacMillan Press Ltd, London

Gordon JW, Scangos GA, Plotkin DJ, Barbosa JA, Ruddle FH (1980) Genetic transformation of mouse embryos by microinjection of purified DNA. Proc Natl Acad Sci U S A 77:7380–7384

Peter Gorer A (1948) The significance of studies with transplanted tumours. Br J Cancer 2:103–107

Grüneberg H (1952) The genetics of the mouse (2nd edn). Martinus Nijhoff, The Hague

Henig RM (2001) The monk in the garden: the lost and found genius of Gregor Mendel, the father of genetics. Mariner Books

Iltis H (1932) Life of mendel. George Allen, London, p 105

Jacob F (1983) Expression of embryonic characters by malignant cells. Ciba Found Symp 96:4–27

Jones EP, Eager HM, Gabriel SI, Jóhannesdóttir F, Searle JB (2013) Genetic tracking of mice and other bioproxies to infer human history. Trends Genet 29:298–308

Keeler CE (1931) The Laboratory Mouse: its origin, heredity, and culture. Harvard University Press, Cambridge, 81 p

Kuehn MR, Bradley A, Robertson EJ, Evans MJ (1987) A potential animal model for Lesch-Nyhan syndrome through introduction of HPRT mutations into mice. Nature 326:295–298

Lyon MF (1961) Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190:372–373
Martin G (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci USA 78:7634–7638
McGrath J, Solter D (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell 37:179–183
Michaux J, Reyes A, Catzeflis F (2001) Evolutionary history of the most speciose mammals: molecular phylogeny of muroid rodents. Mol Biol Evol 18:2017–2031
Mintz B (1962) Formation of genotypically mosaic mouse embryos. Am Zool 2:432
Moriwaki K, Shiroishi T, Yonekawa H (1994) Genetics in wild mice: its application to biomedical research. Japan Scientific Societies Press, Tokyo
Morse HC 3rd (1978) Origins of inbred mice. Academic Press, New York
Morse HC, 3rd (1981) The laboratory mouse—a historical perspective. In: Foster HL, Small JD, Fox JG (eds) The mouse in biomedical research, vol 1. Academic Press, New York, pp. 1–16
Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ (2001) Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618
Musser GG, Carleton MD (1993) Family muridae. In: Wilson DE, Reeder DM (eds) Mammalian species of the world, 2nd edn. Smithsonian Institution Press, Washington, pp 501–755
Paigen K (2003a) One hundred years of mouse genetics: an intellectual history. I. The classical period (1902–1980). Genetics 163:1–7
Paigen K (2003b) One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981–2002). Genetics 163:1227–1235
Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, Sargent EE, Asquith S, Crew V, Johnson KA, Robinson P, Scott VE, Wiles MV (2004) An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 14:1806–1811
Rader K (2004) Making mice: standardizing animals for American Biomedical Research, 1900–1955. Princeton University Press, New Jersey
Rostand J (1957) Un précurseur de Mendel: le pharmacien Coladon. C R Hebd Seances Acad Sci 244:2973–2974
Russell ES (1978) Origins and history of mouse inbred strains: contributions of clarence cook little. In: Morse HC 3rd (ed) Origins of inbred mice. Academic Press, New York, pp 45–68
Sage RD, Atchley WR, Capanna E (1993) House mice as models in systematic biology. Syst Biol 42:523–561
Shimkin MB (1975) A. E. C. Lathrop (1868–1918) mouse woman of Granby. Cancer Res 35:1597–1598
Snell GD (1978) Congenic resistant strains of mice. In: Morse HC 3rd (ed) Origins of inbred mice. Academic Press, New York, pp 119–156
Steensma DP, Kyle RA, Shampo MA (2010) Abbie Lathrop, the "mouse woman of granby": Rodent Fancier and accidental genetics pioneer. Mayo Clin Proc 85:e83. doi:10.4065/mcp.2010.0647
Stevens LC (1970) The development of transplantable teratocarcinomas from intratesticular grafts of pre- and postimplantation mouse embryos. Dev Biol 21:364–382
Stevens LC, Little CC (1954) Spontaneous testicular teratomas in an inbred strain of mice. Proc Natl Acad Sci USA 40:1080–1087
Strong LC (1978) Inbred mice in science. In: Morse HC 3rd (ed) Origins of inbred mice. Academic Press, New York, pp 69–75
Surani MA, Barton SC, Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature 308:548–550
Suzuki H, Aplin KP (2012) Phylogeny and biogeography of the genus Mus in Eurasia. In: Macholán M, Baird SJE, Munclinger P, Piálek L (eds.), Evolution of the house mouse. Cambridge studies in morphology and molecules: new paradigms in evolutionary biology. Cambridge University Press, Cambridge, pp 35–64

Szatkiewicz JP, Beane GL, Ding Y, Hutchins L, Pardo-Manuel de Villena F, Churchill GA (2008) An imputed genotype resource for the laboratory mouse. Mamm Genome 19:199–208

Tarkowski AK (1961) Mouse chimaeras developed from fused eggs. Nature 190:857–860

Wakayama T, Yanagimachi R (1999) Cloning of male mice from adult tail-tip cells. Nat Genet 22:1217–1218

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F (2007) On the subspecific origin of the laboratory mouse. Nat Genet 39:1100–1107

Chapter 2
Basic Concepts of Reproductive Biology and Genetics

2.1 Introduction

This chapter brings together a variety of information and concepts that are important for understanding the following chapters. The first section is an overview concerning mouse reproductive biology and embryology. This topic is important because, nowadays, many experiments in genetics require the manipulation of embryos at different stages of development, either to study their phenotype or for the production of chimeras with other embryos or with genetically engineered embryonic stem (ES) cells. The second part is a compilation of concepts of general or molecular genetics related to the phenotypic expression of mutations. More information can also be retrieved from several websites, where books and manuals are freely available online.[1]

2.2 Reproduction in the Laboratory Mouse

2.2.1 The Estrous Cycle and Pregnancy

Laboratory mice are *polyestrous* mammals. This means that, provided they are raised and housed in a suitable environment, the animals can reproduce all year round with only a small decline in fertility during the winter season.[2] In females, sexual maturity (puberty) takes place gradually from the age of 3–4 weeks. The vaginal orifice, which is normally sealed at birth by an epithelial operculum, opens

[1] The website http://informatics.jax.org/ is a fundamental database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data. It is a true "gold mine" for mouse geneticists to which we will frequently refer. Several books dealing with some fundamental aspects of mouse biology are freely available at this website.

[2] The reproductive activity of wild mice is interrupted or reduced during winter. This period is called *anestrus*.

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_2

between 25 and 40 days. From 6 to 8 weeks after birth, and depending on the strain, ovulation starts, and, in principle, all females older than 8 weeks are able to reproduce, exhibiting a typical cyclic sexual activity. Male puberty occurs slightly earlier, sometimes as early as 5 weeks, usually at 6–8 weeks.

The female reproductive cycle, the *estrous cycle*, lasts 4–6 days and is arbitrarily divided into four stages with the following order: *proestrus*, *estrus*, *metestrus*, and *diestrus*.[3] Proestrus and metestrus last about one day each, while the estrous period lasts only 12–16 h. Diestrus is the last and longest stage of the estrous cycle (~2 days).

Based on vaginal cytology, embryologists have defined criteria that characterize the four stages of the mouse estrous cycle (Byers et al. 2012). According to these criteria the estrus period is characterized by the presence of many flat and keratinized epithelial cells that are obvious upon examination of vaginal swabs. These cells are eosinophilic, meaning that they are stained deep red by the dye eosin. These visible changes during the estrous cycle reflect the variations in progesterone and estrogen levels. Female mice copulate only during the estrous period, which is often designated the "*heat period*" by analogy with the sexual behavior of other domestic females. The heat period lasts about 12 h and mating generally occurs during the first half of the night. In mice, matings are uncommon during the day.[4]

By using the above-described cytological parameters it is possible to identify and sort out the female mice that are in the estrous phase of the cycle, and, accordingly, that are hormonally prepared to copulate. However, this procedure is tedious and labor-intensive, especially when many females are to be selected, and for this reason it is not used very much. In practice, researchers prefer to select the female mice that are in the best conditions to mate by examining the external vaginal morphology (Byers et al. 2012). In this case, the vulva is slightly swelled and the vagina is slightly open. This kind of selection requires some experience but it is fast, quite reliable, and has the enormous advantage of not stressing the mice in a critical period.

The proestrus and estrus phases of the cycle are often designated the *follicular phase* because it is at the end of this phase that a batch of mature oocytes is released from the ovarian follicles. This generally occurs during or immediately after copulation, but copulation is not a prerequisite for this to occur because mice are spontaneous ovulators. If males are not present in the cage, ovulation will still normally occur during estrus.

Shortly after copulation, the fluids secreted by the various sexual glands of the males (in particular the seminal vesicles and the coagulating glands), which are components of the male's ejaculate, coagulate to form a vaginal plug. The plug in

[3] *Estrus*, sometimes spelled *oestrus* (UK), is a noun; *estrous* (*oestrous*) is the corresponding adjective.

[4] For some precisely timed pregnancies, female mice must sometimes be bred in a "light-reversed" environment.

question tightly seals the vaginal lumen and prevents any further mating.[5] The vaginal plug is a relatively hard substance and remains in the female's vagina for several hours (up to 6–8 h or even more). During this time the vaginal plug progressively resorbs and the spermatozoa are released. Detection of a vaginal plug means that mating occurred during the preceding hours, but does not guarantee that pregnancy will ensue.[6]

By analogy with the follicular phase, metestrus and diestrus constitute the *luteal phase*. During this phase the *corpus luteum* forms and replaces the follicle. The *corpora lutea* secrete the hormone progesterone, the hormone of pregnancy, and persist until the end of pregnancy—if pregnancy ensues. If not, the corpora lutea degenerate and a new cycle starts. Corpora lutea are easy to recognize at the surface of the mouse ovary because they are slightly protuberant and often stained light orange. After fixation with formalin or Bouin's fixative, their identification is even easier.

When virgin or non-pregnant females are housed in groups and mated with males without prior selection of the phase of the estrous cycle, the frequency of natural mating is not evenly distributed over the following nights. On the contrary, one generally observes a peak after the third night of mating, indicating that some synchronization of the estrous cycle occurred. Synchronization of the estrous cycle by the presence of a male has been reported and is called the *Whitten effect* (Whitten 1956). It is a consequence of the dispersion in the environment of volatile pheromones that are at high concentration in the urine of males; these pheromones interfere with the hormonal control of the female cycle.

Fertilization of the oocytes takes place 10–15 h after ovulation, in the upper segment of the female reproductive tract, more precisely during their transit through the Fallopian tubes or oviducts (sometimes called *ampulla*). When the head of a sperm cell succeeds in penetrating the oocyte after passing through the *zona pellucida* (also designated *oolemma*), the penetration of other sperm cells is blocked and this triggers the completion of the second meiotic division. The second polar body from the oocyte is ejected within two hours; the male pronucleus expands, and finally the two haploid pronuclei (male and female) fuse, and the oocyte becomes an egg (i.e., a diploid embryo that is not yet implanted). Segmentation in the embryo begins slowly at first. 68–72 h after fertilization (i.e., at the beginning of the 4th day after mating), the embryos enter the uterus and implant into the uterine wall at the late or expanded blastocyst stage.

[5] Such a vaginal plug is specific to the *Mus* genus and does not exist, for example, in the rat. Whether it confers a selective advantage to the species is an open question.

[6] As mentioned, most matings occur during the night; this is why "plugging" must be achieved preferably during the morning of the following day. Detection of a plug is sometimes very easy, especially when it bulges out of the vagina. In other instances, a probe may be necessary to detect resistance when gently inserted into the vagina. The type of probe used by ophthalmologists to unclog the tear ducts of human patients is a perfect tool for this task.

Embryologists date the different stages of pregnancy from the day the vaginal plug is discovered—i.e., day E0.5 by convention.[7]

Starting at 12–14 days of gestation, it is possible to detect the fetuses implanted inside the uterus, which feel like "rosary beads" to the touch. To do this, the female must be held firmly by the skin of its neck and back, with its abdomen overturned, and gently palpated with the fingers of the other hand once the abdominal wall is relaxed. Around 12 days of gestation, the pregnant females start to gain weight and will soon show abdominal bulging; this can be another way to confirm pregnancy by comparison with age-matched non-pregnant females.

Matching the number of corpora lutea with the actual number of fetuses implanted in the uterine horns allows one to compute the number of conceptuses that were possibly lost before implantation. This may be important, for example, when an embryonic lethal mutation is suspected to be responsible for the reduction in the size of the progeny. In normal conditions, the number of corpora lutea, which can be counted directly under a magnifying glass corresponds to the number of implanted fetuses (see Sect. 2.2.7 on twinning).

The gestation period ranges from 19 to 22 days but this depends upon a number of parameters. For example, females that are pregnant for the first time (primiparous) deliver their progeny up to 1 day before multiparous females of the same strain. The duration of pregnancy also varies slightly from one strain to another. For example, pregnancy is, on the average, 1 day longer in mice of strain DBA/2 than in mice of strain C57BL/6.

At the end of the gestation period, the corpora lutea degenerate (*luteolysis*), inducing parturition.[8] The pelvic girdle of the females relaxes and parturition begins in the following 2–4 h.[9] During the same period, the behavior of the female changes dramatically. The female is hyperactive and appears to have only one thing in mind: preparing a nest in a corner of the cage, preferably in a darker area.

Parturition generally occurs at night and may last up to 3 h, depending on the litter size. The fetuses are expelled one after the other, giving the mother time to take care of each of the pups. The fetal membrane and the placenta, as well as the dead embryos, if any, are carefully removed and ingested by the mother.[10] Embryos are also stimulated for breathing by repeated gentle pressure of the mother's paws on the thorax of the newborns. Once the last pup has been delivered and carefully revived, the mother lays over all the newborns gathered in the nest and lactation starts. Newborn mice are hairless, deaf, and blind, and are unable to regulate their body temperature

[7] Dating the different steps of mouse embryonic development has been a matter of controversy. Some embryologists wanted the first day of pregnancy to be designated day 1; others argued that it should be day 0. In fact, the most accurate dating takes into account that, when the vaginal plug is discovered, the embryo is at 0.5 days of development. At this time it is a one-cell embryo just after fertilization (E0.5) (based on Theiler 1972).

[8] Resorption of the corpora lutea is triggered by prostaglandins secreted by the placenta.

[9] A gentle pressure on the pelvis of the mouse allows one to detect the relaxation of the pelvic girdle.

[10] Making the observation of non-viable (stillbirth) phenotypes difficult.

for the first 2 days of life *ab utero*; this is why the mother leaves the nest for only brief periods, only to feed, defecate, and drink. Lactation normally lasts 3–4 weeks depending on the number and degree of vigor of the pups. In the mouse, the number of neonates is frequently greater than the number of nipples (10), but this is not a problem and the pups are generally fed adequately.[11] From the age of 12–14 days, the young mice start eating solid food and the mother's milk is only a complement to the diet. At the end of the lactation period, in general at the end of the third week of life, the young mice are weaned and separated according to their sex by the technicians.

The standard reproductive cycle we have just described is sometimes modified to fit with practical contingences. For example, adoption and foster nursing are common practices in laboratory mouse breeding colonies, especially when the number of progeny is low or the mother is not particularly good at nursing. When there are only one or two pups in a progeny, the mother frequently abandons it/them, presumably because the stimulation of milk production is insufficient. If this situation occurs, it is then wise to take no risk and to transfer the secluded pups as early as possible into an age-matched (up to 1 day younger) litter.[12] Mice dams, unlike many other female mammals, generally accept adopted pups to nurse and milk, especially when they are young. Newborns selected for adoption can be simply added or exchanged in equal numbers with pups of the foster mother. It is recommended, when possible, that the newborns to be adopted be put in contact with some urine-soaked wood-shavings taken from the mother's bedding prior to the transfer, to expose them to the foster mother's smell.

Female mice can deliver up to eight progenies in their sexual life, depending on the strain. However, the progeny size decreases after the fourth progeny and, most importantly, the time that elapses between two successive progenies increases after the third progeny. The number of progeny one can expect from a group of female breeders can be evaluated based on the breeding records.[13] Males can breed for a very long time, sometimes up to 2 years; however, they are normally replaced after 10–12 months, depending on the strain.

Although mice are legendary for having exceptional aptitudes with regard to reproduction in the wild, the situation is different in laboratory conditions and sometimes requires special care. Reproduction and sexual behavior can be influenced by a number of parameters that are not always easy to control. Pheromones, for example, which are true olfactory hormones, play a major role in this matter. The mouse is probably more affected by pheromones than any other mammal, because of the complexity of its olfactory functions. Pheromones are proteins which are released into the urine, the skin secretions, and the saliva of males and

[11] If this is not the case, the pups are left outside of the nest; they progressively cool, do not move much, and have no milk in their stomachs. Foster nursing is then urgent.

[12] Selecting a mother nursing a litter with a different coat color (albino/non-albino) is a clever way to check the success of the adoption without perturbing the mother.

[13] A useful and reliable criterion is the average number of mice weaned per mated female per week.

which modify the behavior of females. We have already reported the *Whitten effect* (synchronization of estrous cycle) that affects female mice when they are housed in groups. In addition to this observation, when females are kept in the absence of male pheromones (which is not easy to achieve in practice), this leads to a state of anestrus (lack of a normal estrus cycle). This phenomenon is called the *Lee–Boot effect* (Van der Lee and Boot 1956). Finally, it is sometimes observed that females, although found with a vaginal plug, never get pregnant when housed in close vicinity with some males. This phenomenon is known as the *Bruce effect* and an explanation is that the pheromones of the males prevent embryo implantation. The males in question are called "*strange males*" (Bruce 1959).

Nutrition is another major parameter that must be seriously taken into account concerning mouse reproduction. Since laboratory animals are fed exclusively on industrial (pelleted) diets, it is extremely important to make sure that the diet constantly provides the optimal amount of nutrients and vitamins, even after sterilization by heat or gamma rays. Some vitamins (C, B1, B9 for example) are extremely heat-sensitive but yet are essential to the function of reproduction; it is therefore essential to frequently change the heat-sterilized food. Nutritional deficiencies are difficult to diagnose but they are insidious and almost always have consequences on fertility, even if the mice do not exhibit any other obvious signs.

Environmental conditions (temperature, ventilation, noise, light cycle) are other parameters to be controlled with care. Noise and vibrations are probably the worst, especially when discontinuous, because the animals cannot become familiar with them and are in constant stress. When the airflow bothers the animals they generally protect themselves and their nest by building a bulwark with their bedding. This is a good indication that something is wrong with the air-conditioning system or the airflow inside the individually ventilated cage. Environmental enrichment like nesting materials and igloos are highly recommended to improve the breeding performance of a mouse colony.

Finally, infectious diseases are also extremely important and must be carefully monitored. Some viruses that cause unapparent diseases have a strong influence on fertility, either because they interfere with the production gametes or because they result in abortions or stillbirths. For more details concerning husbandry and maintenance of laboratory mice, consult the books by Fox et al. (2007) and Hedrich (2012).

2.2.2 Inducing Ovulation in the Mouse (Superovulation)

The information provided in the preceding section concerning mouse reproduction will be useful for those scientists who are willing to run a breeding unit. However, in many cases, geneticists are only interested in harvesting large quantities of fertilized eggs, for example, for creating transgenic animals by pronuclear injection, or for making chimeras, or simply for the preservation of embryos at low temperature. In some other cases, researchers are only willing to collect unfertilized oocytes for in vitro fertilization. In these cases, young females aged 3–5 weeks

(prepubertal females) are treated by injection of gonadotropin hormones and subsequently mated either to fertile or to vasectomized studs depending on the aim of the experiment. In practice, the females in question receive a first injection of 2.0–5.0 international units (IU) of the gonadotropin PMS (pregnant mare serum) in the afternoon of day 1, an injection that artificially induces a first estrus in the young females. 42–50 h later, they receive another injection of 2.0–5.0 IU of the gonadotropin HCG (human chorionic gonadotropin) that artificially induces ovulation.

Responses to gonadotropin injections vary from one strain to another, and, for this reason, the optimum doses and age of the mice to be injected must be determined, for each strain, by doing preliminary experiments[14] (Luo et al. 2011). Ovulation occurs approximately 12 h after the HCG injection, at which time the eggs (fertilized or not) can be collected by flushing the oviducts with a syringe filled with the culture medium. In the best experimental conditions the treated females can produce up to 30–40 embryos (hence the name superovulation), although the response to the hormonal treatment is highly variable between inbred strains (BALB/c is known to be a poor responder, while FVB is a high-responder). Female mice can also be superovulated after puberty, but in this case the production of embryos is much less efficient, presumably because the gonadotropin treatment interferes with the hormonal status of the treated female. If fertilized eggs are to be collected, it is important to mate no more than three or four hormonally treated females per male. The males should be older than 8 weeks and, ideally, proven breeders. Looking for vaginal plugs the following morning is then necessary to select the females that will be sacrificed to collect the embryos (at the desired stage). In order to be ready for the transfer of the manipulated embryos, it is essential to produce pseudopregnant females that will serve as recipients. This is typically achieved by mating outbred females (see Chap. 9) to vasectomized (sterile) males (created through a simple surgery). This mating is necessary for the uterine environment to become receptive, since only pseudo-pregnant females will allow the successful implantation and development of the fostered embryos. For more information and detailed protocols on these techniques, refer to the excellent manual by Nagy et al. (2003) and visit the webpage of the International Society for Transgenic Technologies (ISTT) at http://www.transtechsociety.org/.

2.2.3 Artificial Insemination

Several techniques for artificial insemination (AI) in the mouse have been described in the past (Wolfe 1967; Leckie et al. 1973). These techniques are simple and do not require sophisticated or expensive equipment. The sperm is taken from the vas deferens or the epididymis, mixed at room temperature in a

[14] The response to gonadotropin injections may also vary from one batch of hormone to the next.

few milliliters of tissue culture medium, and injected directly into the uterus of the recipient female (at least 3×10^6 spermatozoa) using an insulin-type syringe, with a blunted needle, and a speculum to avoid harming the vaginal walls.[15] In this case, however, the vasectomized male must not be placed with the female before insemination, because the vaginal plug would interfere with the process. Capacitation of the spermatozoa does not seem to be a problem in this case.

Another technique has been reported where the sperm cells are injected directly into the upper uterine horns or the ampulla with a glass micropipette after laparotomy (uterine insemination) (De Repentigny and Kothary 1996). This second technique does not require such a high number of sperm cells, as compared to vaginal insemination.

Whatever the technique used, the yield in terms of embryo produced per inseminated female is quite low compared to other species. In spite of this low efficiency, artificial insemination has proven useful for obtaining hybrids between laboratory mice and mice of different species of the *Mus* genus (*Mus caroli* or *Mus cervicolor,* for example) because mice of some of these species do not copulate spontaneously with laboratory mice (West et al. 1977). Artificial insemination was also used for studying the possible mechanisms leading to segregation distortion in the progeny of males heterozygous for *t*-haplotypes[16] (Olds-Clarke 1989).

When given a choice, one must remember that F1 hybrids or outbred females have higher levels of fertility when used for AI. In addition, successful insemination can only occur when the inseminated female is in the late proestrus/early estrus stage.

AI will probably not be used very much in the future, because alternative techniques exist that are more reliable and have a much better yield.

2.2.4 In Vitro Fertilization in the Mouse

In vitro fertilization (IVF) is the most frequently used technology for assisted reproduction in humans. The technology was adapted to the mouse several years ago but this has not been easy to achieve and many critical steps had to be overcome (Whittingham 1968; Vergara et al. 1997). A major difficulty has been the development of suitable culture media allowing for a good rate of survival for the early mouse embryos. Another problem has been to optimize the timing of superovulation regimens for the different strains.

---

[15] An ear speculum is an ideal tool. The extremity of a 20-ml glass pipette would also fit perfectly for this purpose.

[16] The *t*-haplotype is a small chromosomal region of chromosome 17 that is highly polymorphic among wild mice of the *Mus m. domesticus* species. Frequently, *t*-haplotypes of wild origin are not transmitted by heterozygous males in compliance with Mendel's laws (i.e., 50:50), but at a much higher frequency (95:5 or even 99:1).

Nowadays, protocols for IVF are available for most of the strains, even though some of them exhibit a higher rate of fertilization than others (Sztein et al. 2000; Nakagata et al. 2014).

The IVF technique generally consists of four steps: (i) young prepubertal females are injected with gonadotropins as described above; (ii) the morning following HCG injection (~8 h after), the oocytes are collected and gently washed; (iii) the oocytes are mixed for 4–6 h in vitro with either fresh or recently thawed frozen spermatozoa; and (iv) after inspection and selection, the fertilized eggs are transferred into a 0.5-day post-coitum (pc) pseudopregnant female. It is recommended to prepare the sperm sample one or two hours before mixing with the oocytes to allow capacitation to occur, although capacitation of mouse spermatozoa does not seem to be as crucial as it is in other mammalian species.

IVF is the technology of choice when it is desirable to rapidly expand a strain (for example, a transgenic line) from a few males that carry a desired or unique genotype, or for maintaining strains with poor breeding performance. IVF has the advantage that it can be performed using frozen or fresh sperm. The technique can also be used for the re-derivation of infected mouse colonies, and is frequently used for the transfer of genotypes of interest between laboratories.

2.2.5 Ovary Transplantation

When a mutant or transgenic female is potentially fertile (i.e., when it produces viable oocytes) but is unable to breed because of some kind of handicap, an ovarian transplantation is a good option. Some classic mutant mice such as *dystrophia muscularis* (*Lama2*^{dy}), *obese* (*Lep*^{ob}), and *dwarf* (*Pou1f1*^{dw}) were historically maintained by performing serial ovary transplantation. The technique consists of the surgical removal of the ovaries of the infertile donor female (even from very young females), and transfer into the ovarian bursa of an ovariectomized recipient female. Again, either freshly collected ovaries from the donor female or stored frozen/thawed ovaries can be used. Use of a recipient female with a different coat color from the donor is recommended to differentiate pups accidentally generated from residual ovary tissue (genotype of the recipient female).[17] Although the ovarian bursa is an immunologically privileged site, it is convenient to use recipient females that are histocompatible with the donor female. Alternatively, immunodeficient females (e.g., *nude* and SCID mutants) can be used as recipients.

[17] It is for the rapid and safe identification of the origin of its progeny that mice of the strain 129/J segregate for the coat color alleles *Tyr*^{c} and *Tyr*^{ch}.

2.2.6 Intra Cytoplasmic Sperm Injection

Intra cytoplasmic sperm injection (ICSI, also known as *micro-insemination*) is another technology that is commonly used in humans to overcome persistent male infertility problems (for example, oligospermia, teratozoospermia, incapacity of the spermatozoa to pass through the zona pellucida, etc.). Here again, the technology has been adapted to the mouse with roughly the same basic protocol as in humans. In short, mature oocytes are held at the tip of a micropipette, by gentle suction, while a sperm head is injected deep into the cytoplasm of the oocyte by using a piezo-driven micromanipulator. This equipment and procedure allow for the safe injection of sperm heads by making only a very small hole in the zona pellucida that is promptly resealed once the needle is withdrawn (Ogura et al. 2003; Ogonuki et al. 2011). After the procedure, the oocyte is placed into an appropriate culture medium where its development is checked for a few hours.

ICSI has been adapted for use with immature (haploid) spermatogenic cells (round spermatids or elongated spermatids), and high rates of offspring development have been obtained (~30 % in some cases).

ICSI and ROSI (round spermatid injection) technologies have also demonstrated some practical advantages in the mouse. ICSI, for example, allowed for the recovery of normal pups from spermatozoa taken from the testes or epididymides of dead mice whose bodies had been stored at low temperature (between −20 °C and −80 °C) for a few years (Ogura et al. 2005; Ogonuki et al. 2006).

ROSI technology has also been cleverly used to reduce the time required for the development of fully congenic mouse strains by using the nucleus of round spermatids removed from young males (22–25 days of age) for the fertilization of superovulated oocytes flushed from 3-week-old females. With this technology, a backcross generation could be reduced to only 41–44 days, and a fully congenic strain (homozygous for 97.7 % of 176 tested markers) could be produced in 190 days (~6 months) (see Chap. 9) (Ogonuki et al. 2009).

2.2.7 Cryopreservation of Mouse Embryos and Spermatozoa

The mouse was the first mammal whose embryos were successfully frozen and stored at very low temperature. The methodology, which was published in 1972 (Whittingham et al. 1972; Wilmut 1972), required slow cooling (0.3–2 °C/min) and slow warming at some critical steps as well as the use of cryoprotectants to prevent ice crystals from damaging the cells of the embryo. In these initial experiments the cryoprotectants were either dimethyl sulfoxide (DMSO) or glycol. Since these pioneering experiments, the technique has been improved and nowadays mouse embryos are routinely stored at very low temperatures (in liquid nitrogen at –196 °C) for virtually unlimited periods and thawed when

requested with quite high rates of survival.[18] Embryo freezing and banking is achieved routinely in many laboratories, and is also available as a service from several commercial institutions. Short courses and demos with tutorials are available in several formats, for example as "webinars" or highly didactic movies, and are freely available through the internet.

Vitrification is another method of cryopreservation that has been developed more recently. With this method the embryos are osmotically dehydrated and then cooled by a rapid transfer into liquid nitrogen.

Cryopreservation of mouse spermatozoa has proved capricious for a long time and its rate of success is still relatively strain-dependent; for example, C57BL/6 sperm is difficult to freeze and the proportion of unviable sperm cells after thawing is quite high. However, the technology is rapidly improving and it is likely that most of the technical problems that still remain nowadays will be adequately solved in the near future (Sztein et al. 2000; Nishizono et al. 2004; Nakagata et al. 2014).

Freezing embryos and spermatozoa both represent a safe and (relatively) cheap way of exchanging mouse strains between different laboratories across the world. This practice has the advantage of reducing the risk of transmission of infectious diseases, a great concern for most veterinarians in charge of laboratory animal facilities.

Ovarian cryopreservation has been demonstrated to be another valid option for banking mouse genetic resources; in particular, it is the only technique that can be used to preserve oocytes from aged or problematic female breeders (Sztein et al. 2010).

Readers who are interested in the practice of cryopreservation technologies can refer to comprehensive reviews on the subject by highly experienced authors (Glenister and Rall 2000; Sztein et al. 2010; Nakagata 2011; Mochida et al. 2011). A didactic movie is also freely available on the internet: see reference list.

2.2.8 Twinning in the Mouse

The existence of the spontaneous occurrence of identical twins in the mouse is still debated. According to Grüneberg (1952), twinning occurs in the mouse as in many other mammalian species, but extremely infrequently; and twins may experience a disadvantage during their early embryonic life. Identical twins have been

[18] Experiments performed at the Harwell (MRC) Research Centre have demonstrated that the damage caused by radiation (cosmic rays) to mouse embryos when stored at low temperatures for very long periods is practically negligible.

occasionally observed in utero at very low frequency, between embryonic days 8 and 10, but such embryos have not been recorded by embryonic day 16–17.[19, 20, 21]

McLaren and colleagues, looking for identity by DNA fingerprinting (using human minisatellite probes) in litters segregating for ten genetic loci, did not find any evidence of twinning in a population of 2,000 outbred mice. The authors concluded that twins are either extremely rare in the stock of mice they studied, or that they have such reduced viability that their chance of surviving to weaning is low (McLaren et al. 1995).

Spontaneous twinning is uncommon in the mouse; however, the experimental production of monozygotic twins by embryo splitting has been successfully achieved in several laboratories. Illmensee and colleagues demonstrated that in vitro splitting of mouse embryos at the 2-, 4- and 8-cell stage, followed by their transfer into empty zonae pellucidae, could be achieved with a relatively high rate of success. Embryonic development was monitored after in vitro culture for a few days and twin blastocysts from 2- and 4-cell splitting showed well-developed colonies with trophoblastic cells and clusters of inner cell mass (ICM) cells (Kaufman and O'Shea 1978; Illmensee et al. 2005).

2.2.9 Cloning Laboratory Mice

Cloning is an asexual method of reproduction that is commonly used in plants (e.g., cutting or striking) as well as in some insects: it offers the possibility of obtaining a potentially unlimited number of genetically identical individuals. In mammals, clones have also been produced experimentally by embryo splitting. More recently, cloning has been achieved by the experimental replacement of the nucleus of an unfertilized oocyte by the nucleus of a specific somatic cell from the same species, a process known as somatic cell nuclear transfer (SCNT). In most species, these experiments have been very difficult to perform, with low rates of success. Beyond these difficulties, many clones have developed severe pathologies that in many cases have undermined the interest of the enterprise. Cloning is no easy endeavor and many fundamental questions regarding possible modifications at the genome level during the early stages of development still remain to be understood.

---

[19] It is not easy to observe twins by the mere examination of the implants in the mouse uterus, as placental fusion is frequent in this species.

[20] Discordances between the number of implants (dead or alive) and the number of corpora lutea does not support the idea that twinning commonly occurs in the mouse.

[21] Twinning (sometimes called “polyembryony”) is the rule in nine-banded armadillos of the South American species *Dasypus novemcinctus*. In this species, the females regularly deliver progenies composed of four monozygotic twins. This regular production of genetically identical offspring makes the species a valuable model for multiple births.

Cloning the laboratory mouse has also been relatively difficult to achieve for technical reasons. Nonetheless, cloned mice were produced for the first time after the transplantation of nuclei taken from cells of the *cumulus oophorus*, hence the name of the first cloned female mouse: "*Cumulina*" (Wakayama et al. 1998).[22] Since then, mice have been cloned from a variety of different donor cells, including fibroblasts (tail skin), olfactory sensory neurons, ES cells, bone marrow cells, and liver cells. Recently, live mice have also been obtained after transplantation of the nucleus of peripheral blood leukocytes into enucleated oocytes from a drop of blood (Kamimura et al. 2013). Mice cloned from cumulus cell nuclei have even been themselves cloned in series for 25 generations, producing over 500 viable, fertile, and healthy clones derived from the original (single) donor. These experiments proved that serial recloning over multiple generations is possible in the mouse (Wakayama et al. 2013).

Compared to the situation in other species, in particular domestic species, the cloning of mice has relatively limited applications. This is because it is very easy in this species to obtain large populations of mice with exactly the same genotype. For example, mice of an inbred strain or born from a cross between two inbred strains are all genetically alike exactly as if they were cloned individuals (same genotypes). In these conditions, cloning mice may only help to enhance our understanding of the technical and biological factors that contribute to successful cloning in a species of economical interest. Experimenting with mice, biologists may be able to understand how the donor nucleus taken from a differentiated cell becomes reprogrammed by the oocyte cytoplasm to enable it to give rise to the different cell types. Similarly, the cloning of mice may help in the understanding of the reversibility of epigenetic changes occurring during tissue differentiation.

2.2.10 Mosaics and Chimeras

The terms *mosaic* and *chimera* are frequently incorrectly used in the scientific literature, even under the signature of professional geneticists. Mosaics are organisms composed of cells with a different genetic constitution, although deriving from one single conceptus (embryo). For example, an organism composed of cells with a different karyotype is a typical mosaic when this results from the loss or abnormal disjunction of a chromosome during the many mitoses that occur throughout embryonic development. An abnormal disjunction generates daughter cells with $2n - 1$ chromosomes and cells with $2n + 1$ chromosomes, and these cells are themselves mixed with normal 2n cells in variable proportions.[23] Such "chromosomal mosaics" are often viable, especially if the mosaicism concerns the

[22] Cells of the *cumulus oophorus* are ovarian (but somatic) cells. They surround the oocyte and are shed with it upon ovulation.

[23] Cells with $2n + 1$ chromosomes (trisomic) are in general more viable than cells with $2n - 1$ chromosomes (monosomic).

X-chromosome or a minority of cells (see Chap. 3). 2n/3n and 2n/4n *mixoploid* mosaics have also been described in several mammalian species, including the mouse.

Mosaic organisms composed of normal cells and cells carrying a point mutation at a specific locus are probably very common (this point will be discussed in Chap. 7), but this mosaicism remains unnoticed if it has no deleterious consequences for the mutant cell. On the contrary, when spontaneous mutations accumulate in a cell (or group of cells) that provide the cell with the potential to divide indefinitely or to resist cell death, then these cells may become cancerous (malignant). In this sense, a mammalian organism affected by a cancer can be considered as a true mosaic since the malignant cells have indeed acquired a genetic constitution different from the non-malignant ones although they share the same origin.

Somatic (or mitotic) crossing-overs are yet another way of generating mosaic organisms, but only very few cases have been reported and documented in the mouse (Panthier et al. 1990). To conclude the definition of a mosaic, we note that, in general, mosaics are produced naturally, with no human intervention, while this is not the case with chimeras.

Chimeras (or chimaeras) are organisms that are composed of two (or more) different populations of genetically distinct cells (originated from different embryos), which generally result from human intervention. For example, an immunodeficient mouse that survives because it has received allogeneic bone marrow transplantation is a chimera, as is a mouse that results from the in vitro fusion of two or more morulae of different genetic origins (for example, from two different inbred strains). In this chapter, we will consider exclusively the case of chimeras resulting in a single complete mouse organism with pluri- or multiparental origin.

Mouse chimeras were created almost simultaneously in the early 1960s by Mintz, working at the Fox Chase Cancer Institute (Philadelphia, USA) and by Tarkowski, working at the University of Warsaw (Poland) (Tarkowski 1961, 1998; Mintz 1962). The first chimeric mice were produced by merging two independent morulae in vitro whose zona pellucida (oolemma) had been previously removed by a brief treatment with the enzyme pronase. These chimeras are referred to as *aggregation* or *allophenic* chimeras.[24] They developed in a chimeric animal of normal size, easily recognizable by a dappled coat color if the parental strains were appropriately selected (Mintz and Silvers 1967).

The aggregation technique developed by Mintz and Tarkowski was replaced in the early 1970s by a microsurgical technique that consisted of the injection of embryonic cells directly into the cavity of blastocysts (Gardner 1971).[25] This technique was later modified and improved in several ways (Brinster 1974; Mintz and Illmensee 1975; Papaioannou et al. 1975; Bradley et al. 1984; Stewart et al. 1994).

[24] Sometimes called *tetraparental* chimeras.

[25] This cavity is often called a *blastocoel*.

Chimeric mice have been and still are important tools in biological research, as they allow us to answer questions related to cell lineage and cell potential with regard to tissue differentiation. By studying the muscles of chimeric mice constructed from two partner strains with different isocitrate dehydrogenase alleles (*Idh1*^{a} and *Idh1*^{b}), it was demonstrated that the in vivo origin of the muscular syncytium is from myoblast fusion and not from repeated nuclear division in a non-dividing cell body (Mintz and Baker 1967).

Studying a series of hepatomas, which occurred in C3H/He × C57BL/6 chimeric mice, researchers found that most of these tumors were derived from cells of the hepatoma-susceptible C3H/He strain. However, they also observed that rare hepatomas were derived from cells of both genotypes, suggesting that some intercellular transmission of tumor information may have occurred or that the transformation might have occurred concurrently in two or more cells, indicating that hepatomas may therefore be genetically complex entities (Condamine et al. 1971).

Nowadays, chimeric embryos are produced routinely by injecting totipotent embryonic cells of different types (for example, embryonic cells from another embryo, embryonic stem (ES) cells that may or may not be genetically engineered, embryonic germ (EG) cells, etc.) into the blastocoel of recipient embryos. After this injection, the cells of the ICM of the recipient embryo merge with the transplanted cells and a chimeric embryo eventually develops to term. Today, the technique is mostly used for introducing a new genotype (that of the engineered ES cells) into the germ line of a chimera, allowing it to be ultimately materialized in a living mouse.

Another technique consists of using tetraploid embryos (which are artificially made by electrofusion of two 2-cell diploid embryos) as recipients for the engineered ES cells. It has been observed that, in this case, only the diploid cells (the ES cells) contribute to the formation of the neonates' body, while the cells derived from the tetraploid embryo will exclusively give rise to the trophectoderm and primitive endoderm. This technique is known as tetraploid complementation and, although not used extensively, it has been successfully used to create mice entirely derived from induced pluripotent stem cells (iPSCs) (Kang et al. 2009).

Another very clever technique resulting in 100 % germline transmission from competent injected ESCs has been developed. This technique consists of using a F1 host embryo (designated the "perfect host" or PH) which selectively ablates its own germ cells via tissue-specific induction of diphtheria toxin. This approach allows competent microinjected ES cells to fully dominate the germline, eliminating competition for this critical niche in the developing and adult animal (Taft et al. 2013).

Although chimeras can be either male or female, in experience the majority is male because most of the ES cell lines are XY. Having male chimeras is actually good because they generally have good germline transmission (Nagy et al. 2003). Tetraparental chimeras can breed if the two embryos at the origin of the chimera are both of the same sex. If this is not the case, for example if one set of cells is genetically female and the other genetically male, intersexuality (and sterility) often results. Even when the two embryos that participate in the formation of the chimera are of the same sex, the fertility sometimes depends on which

cell line gave rise to the ovaries or testes. For this reason, the association of a tetraploid (4n) partner with a diploid (2n) one, as explained above, is particularly advantageous.

The production of allophenic chimeras has been used in various contexts to answer biological questions that would not have been easily answered otherwise. For example, chimeras have been produced to transmit lethal genes in the mouse and to demonstrate allelism of two X-linked male-lethal genes, *jp* and *msd* (Eicher and Hoppe 1973). In another example, viable aggregation chimeras have been made by merging normal embryos with embryo homozygous for the recessive lethal mutation Hairy ears (*Eh*-Chr 15), which indicated that the mutation in question was not cell-lethal (we now know that it is a large deletion) (Guénet and Babinet 1978). Finally, especially noteworthy is the production, by Kobayashi et al. (2010), of the first viable rat–mouse chimeras. In this report, the authors also demonstrated that rat iPS cells could rescue organ deficiency in mice, opening new frontiers for tissue engineering.

2.3 Basic Notions of Genetics

2.3.1 Genes and Alleles

In his famous note reporting the results of his *Experiments on Plant Hybridization* (1866), Mendel alluded to "*factors*" or "*units of inheritance*" that are transmitted from one generation to the next and determine the phenotypic characteristics of plants. Using such words, it is clear that Mendel was referring to the concept of genes, but he did not coin any specific word to define these "*units of inheritance*". Several years later, in 1889, de Vries published a book entitled *Intracellular Pangenesis* in which he led an interesting discussion concerning the "*units*" or "*particle bearers of hereditary characters*", and he recommended that the word pangens be used to specify these particles (de Vries 1910). Finally, it was the Danish biologist Johannsen who proposed, in 1909, that the (Danish) word "*gen*" be used to describe the units of heredity. The same Johannsen also introduced the terms *phenotype* and *genotype,* and almost at the same time Bateson proposed the term *genetics* to describe the science dealing with *gens* (*genes*).

Shortly after the confirmation that DNA was the molecular basis for inheritance (seminal work published by Avery, McCarty, and MacLeod in 1944), the definition of the gene was translated into molecular terms and became "a segment of DNA of variable size encoding an enzyme". This was in compliance with the famous "one-gene-one-enzyme" theory formulated by Beadle and Tatum.[26] This definition was reconsidered when it was recognized that some proteins are not enzymes. The

[26] G.W. Beadle and E.L. Tatum were awarded the Nobel Prize in Physiology or Medicine in 1958 for their discovery of the "*role of genes in regulating biochemical events within cells*".

motto was then changed to "one-gene-one-polypeptide" and the definition of the gene was modified accordingly.

In 2002, once the sequencing of the mouse nuclear DNA was completed, followed by the extensive analysis of the transcriptome[27] and the confirmation that a great number of genes were not translated into polypeptides, the definition of the gene changed again. Nowadays a gene corresponds to a segment of DNA that is transcribed into RNA. Some of these RNA molecules, the messenger RNAs (mRNAs), are translated into polypeptides while many others are not translated but have nevertheless important functions as RNAs (see Chap. 5).

Recently, information collected from the systematic analysis of a single transcriptome revealed that mammalian DNA is pervasively transcribed from both strands, and that the proportion of DNA transcribed into RNAs is much greater than expected. The same analysis also revealed that mammalian genes are not always clearly individualized in the DNA strands; on the contrary, their limits are often difficult to define, with some small genes being nested into larger ones, inserted for example in the introns. Thus, it is seems clear that the concept of the gene will have to be reconsidered and its definition reformulated in the near future. For the time being, we will stay with the idea that a gene is a functional unit materialized in a short segment of DNA, which is transcribed into RNA, and whose inheritance can be followed experimentally generation after generation.

For decades, the *genome* was known as the collection of genes of a given species. The word was coined at the beginning of the twentieth century, and at that time it was used to exclusively define the collection of genes. Nowadays, the concept includes both the genes (i.e., the coding sequences) and the sequences of DNA that are intermingled with the genes and are themselves heterogeneous. Thus, when they refer to the mouse genome sequence, geneticists in fact refer to the sequence of the whole nuclear DNA.

The number of genes in the mouse genome is expected to be in the range of 25,000–30,000 but, for reasons that will be discussed further, this assessment is not accurate and will probably never be. For some genes, the number of copies in the genome varies across the different strains, or even individuals, and many among these genes are non-functional (see Chap. 5 regarding CNVs and pseudogenes). It is also known that some genes are present in some strains (or species) and absent in others. All these variations, of course, hamper the accurate evaluation of the number of genes.

In addition to these inter-strain quantitative variations in the number of genes, we know that several different RNAs (coding and non-coding) can be transcribed from the same gene by the mechanism of alternative splicing (detailed in Chap. 5), and this tremendously increases the number and diversity of the molecules potentially encoded in the genome. Obviously, it is the inventory and classification of all these transcripts that would be important to make, rather than an accurate

[27] The transcriptome corresponds to the full set of RNA molecules that are transcribed from the genome. This point will be extensively discussed in Chap. 5.

assessment of the number of genes. This goal is certainly in the minds of many geneticists, but it is a serious challenge and is difficult to achieve.

Whatever the actual number of genes in the mouse genome, once a gene is biologically defined either in terms of function or structure, it can be precisely localized on a specific chromosome of the species using a variety of techniques. The position of such a gene defines its *locus* (plural *loci*, the Latin word for "place") and we will extensively discuss the strategies used for the localization of the genes in Chap. 4.

Many genes exist in several versions (variants) called *alleles*. The word "allele" is an abbreviation of the ancient word *allelomorph*, which was used in the past to describe the different forms of a gene, detected as different phenotypes. Formerly, the concept of alleles was tightly associated with the concept of mutation producing a phenotypic variant different from the wild type (i.e., the version most commonly found in wild animals). In this case, the new version of the gene was defined as a *mutant allele* and was identified in mice, for example, by a different coat color, a heritable skeletal defect, or a debilitating neurological disease.

The concept of the allele has also progressively changed and nowadays one can say that any change at the DNA level that translates into a phenotype different from the previously known phenotypes defines a new allele, regardless of whether the phenotype associated with the new allele is deleterious. The substitution of a nucleotide in a coding sequence that leads to a change in the global electrical charge of a protein characterizes a new allele because, even if the function of the protein is not affected, one can distinguish by electrophoresis the new protein from the other proteins encoded by the same gene: it is a different phenotype.[28] If the nucleotide substitution modifies the activity of the protein, with deleterious consequences, in this case the new allele is either a *hypomorphic* or *null allele* (see Chap. 7).

Other types of structural variations at the DNA level (for example, the so-called single nucleotide polymorphisms or SNPs) can also be used to distinguish allelic variants (DNA variants in this case), even if these allelic variants do not confer any phenotypic change on the animal. In these conditions the reader may appreciate how the definition of the word allele has evolved with time. In the past, the function of the protein, assessed by its effect on the phenotype of the animal, was crucial to define a new allele. Nowadays, any structural change that can differentiate a gene from another at the same locus defines an allele, regardless of the phenotypic consequences. We will come back to this point when discussing the genetic markers used for gene mapping (Chap. 4).

According to the Mouse Genome Database (as of November 2014), 10,425 genes of the mouse have at least a mutant allele and the mouse genome comprises 40,713 alleles altogether. The whole collection of alleles that are segregating in a given population represent what geneticists call the *genetic polymorphism*. This

[28] The word *electromorph* has been coined to define the alleles characterized by a different global electrical charge.

notion of polymorphism applies to the series of alleles at a specific locus or to all loci of a strain or species.

In the mouse, the gene encoding tyrosinase (*Tyr*), an enzyme that is instrumental in the synthesis of the pigment melanin, was one of the first (if not the first) to be identified based on a variation in coat color. At the *Tyr* locus, one allele encodes a normal, functional tyrosinase, and the other encodes a non-functional enzyme resulting in albinism. Nowadays, over 120 different mutations have been collected at the *Tyr* locus, some of them having a phenotype affecting coat color (for example, chinchilla-$Tyr^{c\text{-}ch}$, extreme-dilution-$Tyr^{c\text{-}e}$, hymalaya-$Tyr^{c\text{-}h}$, to cite a few). However, it is likely that sequencing will identify many others that are not yet detected because they have no obvious phenotype. Such a collection of a series of alleles always represents an interesting resource for geneticists.

The Mouse Genome Database specifies rules and guidelines for mouse and rat gene nomenclature (http://www.informatics.jax.org/mgihome/nomen/gene.shtml), with which it is extremely important to comply because genetic nomenclature is a universal language. We recommend that readers frequently refer to these guidelines, which are presented in a very didactic form with many different examples. In case of doubt, information may also be requested directly from the staff of curators, as explained on the website.

2.3.2 Allelic Interactions

When the alleles at a given locus are the same on both chromosomes, the mouse is *homozygous* and the phenotype that characterizes the allele in question is fully expressed: the situation is simple. When the two alleles are different, the mouse is *heterozygous* and the phenotype depends upon the interactions between the two alleles. To illustrate the situation, we will again consider a gene we are already familiar with: the gene encoding tyrosinase (*Tyr*-Chr 7). As we already mentioned, this gene has several alleles, among which some are non-functional; this is the case with Tyr^{c}. When a mouse has the Tyr^{c} allele on both chromosomes 7 (homozygous), it is albino. In contrast, when the mouse has a non-functional allele on one chromosome 7 and a functional allele on the other chromosome, it is heterozygous and is pigmented just like a wild mouse. The Tyr^{c} allele is said to be *recessive* and the normal allele, or wild-type allele (Tyr^{+} or sometimes only +), is *dominant*. In this case, the lack of functional tyrosinase is completely compensated for at the cellular (melanocyte) level by the presence of a single copy of the normal (wild-type) allele.[29]

[29] When an allele is fully dominant, geneticists often write the genotype *Mut/–*, indicating that the allele in question completely determines the phenotype.

Some other alleles at the *Tyr* locus have less dramatic effects than Tyr^{c} on the synthesis of the pigment melanin and in many cases the mice are pigmented, although always less than or differently from the wild type. For example, mice homozygous for the extreme dilution $Tyr^{c\text{-}e}$ allele appear "*slightly stained or dirty black-eyed white*" (Detlefsen 1921). They have a light grey coat color, almost white, but their eyes are solid black, unlike albino mice. Mice homozygous for the chinchilla allele $Tyr^{c\text{-}ch}$ have a diluted coat color (they really look like chinchillas—hence the name of the mutant allele) but their coat color is much darker than mice homozygous for $Tyr^{c\text{-}e}$. Finally, mice homozygous for the Himalayan allele $Tyr^{c\text{-}h}/Tyr^{c\text{-}h}$ have a remarkable pattern of pigmentation with a mainly white body and light-ruby eyes and only the tip of the nose, tip of the ears, and the tail are normally pigmented (black). This is because the tyrosinase encoded by the $Tyr^{c\text{-}h}$ allele is active only in the colder parts of the body, where the temperature is below 35 °C (the enzyme is heat-labile or thermo-labile). This is the same phenotype observed in Siamese cats.

With so many alleles at our disposition, we could breed a wide variety of mice heterozygous or homozygous for different alleles and we would then discover that the normal allele (Tyr^{+}) is dominant over all other alleles. However, if we grade the phenotypes of the mice based on the decreasing intensity of the coat color for all the possible combinations of the four alleles at the *Tyr* locus- Tyr^{+}; $Tyr^{c\text{-}ch}$; $Tyr^{c\text{-}e}$ and Tyr^{c} we observe that they display an almost continuous gradient of pigmentation from type to albino (i.e. $Tyr^{+}/- > Tyr^{c\text{-}ch}/Tyr^{c\text{-}ch} > Tyr^{c\text{-}ch}/Tyr^{c\text{-}e} > Tyr^{c\text{-}ch}/Tyr^{c} > Tyr^{c\text{-}e}/Tyr^{c\text{-}e} > Tyr^{c\text{-}e}/Tyr^{c} > Tyr^{c}/Tyr^{c}$) (from Silvers 1979). The observation of intermediate phenotypes such as $Tyr^{c\text{-}ch}/Tyr^{c\text{-}e}$ or $Tyr^{c\text{-}e}/Tyr^{c}$ allows for the definition of another kind of allelic interaction that is called *incomplete dominance* or *intermediate dominance*, or sometimes *partial dominance*. In these cases one allele is not completely dominant over another, and the expressed physical trait is in between the dominant and recessive phenotypes. In this context, the phenotype of mice homozygous for the Himalayan allele $Tyr^{c\text{-}h}$ cannot be considered as "intermediate"; they are simply different and unique.

The series of alleles that we described at the *Tyr* locus is common in plants and vertebrate species, and many other examples are available in the mouse. As we already said, in most cases the wild-type allele, the one that is most frequently found in wild mice, is often dominant over all other alleles at the same locus; but this is far from being a rule. At the Agouti locus (*A*-Chr 2), where there is another long series of alleles (over 400) affecting coat color, the wild-type allele agouti (*A*) has an intermediate position: it is dominant over some alleles like black-and-tan (a^{t}), non-agouti (*a*), or extreme non-agouti (a^{e}), but it is recessive to yellow (A^{y}), viable yellow (A^{vy}) and a few other *A* alleles. By the way, it is interesting to note that the yellow allele (A^{y}) in question, although dominant over *A* if we consider the coat color, is nevertheless a recessive lethal when homozygous (see Fig. 1.1). A^{y}/A mice have a beautiful yellow coat color but A^{y}/A^{y} embryos display characteristic

abnormalities at the blastocyst stage and die on the sixth day of gestation.[30] This observation means that the notion of dominance and recessivity must be considered only in the context of a specific phenotype.

True dominant mutations, i.e., mutations for which the phenotype of the heterozygote (*Mut*/+) is indistinguishable from the phenotype of the homozygous mutant (*Mut*/*Mut*), are rare in the mouse and in mammals in general. In most instances, the dominant alleles behave just like the yellow (A^y) allele and are lethal when homozygous. Among the few exceptions are some keratin mutant alleles such as Rex ($Krt25^{Re}$), Caracul ($Krt71^{Ca}$), and possibly a few others such as the coat color mutation Sombre ($Mc1r^{E\text{-}so}$) and the neurological mutation Trembler ($Pmp22^{Tr}$).

Another type of allelic interaction that is extremely common in mammals is *co-dominance*. Co-dominance is when the two alleles at a given locus are both expressed in the phenotype of the heterozygote, which has a phenotype of its own. In most genetics textbooks the concept of co-dominance is exemplified by the AB blood groups in humans, where the AB heterozygotes have a phenotype in which both the A and B antigens are expressed on the red blood cells. Blood groups homologous to the human AB system do not exist in the mouse, but practically all the genes expressed in the form of proteins with different electric charges are co-dominantly expressed. Glucose-6-phosphate isomerase (symbol *Gpi1*-Chr 7) is an enzyme that is expressed in most cells; it catalyzes the conversion of glucose-6-phosphate into fructose-6-phosphate. Several alleles at the *Gpi1* locus have been characterized, of which four are common, viable and functional: $Gpi1^a$ and $Gpi1^b$ are found in laboratory inbred strains, $Gpi1^c$ is a spontaneous mutation of recent occurrence in the BALB/c inbred strain, and $Gpi1^d$ was discovered in wild mice. It is likely that many more alleles (electrophoretic variants) exist in wild mice and have not (yet) been identified. All these alleles are co-expressed in mice heterozygous at the *Gpi1* locus.

When the phenotypes of the different alleles at a given locus are carefully analyzed, interesting observations can be made concerning the allelic interactions. A good example is the case of the locus encoding the enzyme argininosuccinate synthetase (ASS). At this locus, several mutant alleles have been identified in the mouse that are potentially interesting models for the human disease citrullinemia type I (CTLN1, OMIM# 215700). Among all the hypomorphic alleles, two are more interesting than others: $Ass1^{bar}$ and $Ass1^{fold}$, because they faithfully replicate the pathology observed in human patients suffering from CTLN1, with variations in terms of survival rate, developmental delay, and neurological phenotype. Homozygous and compound heterozygous combinations of the two alleles create

[30] These yellow mice posed a problem to Cuénot while he was trying to demonstrate that Mendel's laws also apply to mammals. When intercrossing A^y/A mice, he did not find the expected 1:2:1 proportions of phenotypes for a single gene with two alleles, but instead found a 1:2:0 ratio. However, Cuénot provided the correct explanation for these "unusual" proportions.

a spectrum of severe ($Ass1^{bar}/Ass1^{bar}$), intermediate ($Ass1^{fold}/Ass1^{fold}$), and mild ($Ass1^{bar}/Ass1^{fold}$) phenotypes. However, the observation that the compound heterozygotes, carrying one severe allele ($Ass1^{bar}$) and one mild allele ($Ass1^{fold}$), exhibited a milder phenotype (including residual activity of liver ASS and less pronounced plasma ammonia levels) than mice carrying two copies of the mild allele ($Ass1^{fold}/Ass1^{fold}$) was quite unexpected. Obviously, this warrants further investigation concerning the molecular interactions between the different ASS1 mutant proteins (Perez et al. 2010).

Dominance, recessivity, and co-dominance are the most common forms of allelic interactions but there are also a few others that deserve, at least, a short comment. *Semi-dominance* has been used to characterize mutant alleles where the phenotype of heterozygotes is different (and often intermediate) from both kinds of homozygotes. A typical example is the $Kit^{W\text{-}f}$ allele at the *Kit* locus (Chr 5), where $Kit^{W\text{-}f}/+$ heterozygous mice have a light grey coat with a spot on the belly and a small blaze on the forehead, while heterozygous $Kit^{W\text{-}f}/Kit^{W\text{-}f}$ mice are extensively spotted (Guénet et al. 1979). Amazingly, the tails of these mice perfectly characterize the situation: the tail is normally (i.e., completely) pigmented from the base to the tip in wild-type mice; it is half pigmented in heterozygotes and it is completely unpigmented (white) in homozygotes. Another example is the semi-dominant spontaneous mutation called *Naked* (*N*) on distal chromosome 15 (Hogan et al. 1995). The semi-dominant allelic expression is common in the mouse but it is sometimes used for alleles that would be best characterized as incompletely dominant.

Overdominance is a rather rare condition in which the heterozygotes (*M*/*m*) have a phenotype that is more pronounced than that of either homozygote (*M*/*M* and *m*/*m*). We report such a case of overdominance in Chap. 6 with the *Callypyge* mutation in sheep. No similar mutation has ever been reported in the mouse, but some may exist. Overdominance is sometimes used as an alternative word for *superdominance*. Superdominance characterizes a situation where the heterozygotes have a selective advantage over homozygotes. This is the case, for example, with the human allele that encodes sickle-cell anemia (HBB^s) or drepanocytosis. In the countries where malaria is endemic the gene encoding the defective hemoglobin, although lethal when homozygous, is present in over 40 % of the population, while we would expect it to be strongly counter-selected. This is because the HBB^s allele confers resistance to malaria in the heterozygotes. Homozygotes with the normal allele get sick and sometimes die when infected by *Plasmodium*; homozygotes for the mutant allele also get sick from drepanocytosis but the heterozygotes survive *Plasmodium* infection and do not develop severe anemia. This selective advantage of a specific combination of alleles is probably also found in wild mouse populations but, to date, it has never been described in any laboratory mouse or rat population.

In this review of the different forms of allelic interactions, one must not forget the case of genes that are X-linked. In mammalian species, the males have only one X-chromosome and, in these conditions, the individuals are *hemizygous* for all the genes carried by this chromosome and all are fully expressed. In females, the situation is more complex and the situation will be discussed in full detail in

Chap. 6. Without going into detail, one can say that due to the phenomenon of X-inactivation, which is a mechanism of dosage compensation operating in female mammals, most X-linked genes are functionally haploid and only one copy of every gene is transcribed, while the other copy is switched off. The inactivation of one allele over the other is, in most cases, a random process. In the mouse, a few genes are in the so-called pseudo-autosomal region of the X-chromosome and behave as autosomal genes. The gene encoding steroid sulfatase (*Sts*) is one example.

When a mutation occurs in a mouse population, the allelic interactions exhibited by the novel allele is important information to take into account in the process of genome annotation. If the novel allele is dominant or semi-dominant, it makes sense to guess that the observed phenotype is the consequence of a structural defect of the protein encoded by the mutant allele. On the contrary, when the novel allele is fully recessive, this would indicate a loss-of-function (or hypomorphic) mutation for the protein encoded by the mutant allele. For example, mutations in the genes encoding collagens or fibrillins, which generate a structural defect in the proteins in question, are almost always dominant or semi-dominant.[31] On the contrary, mutations that cause an "*inborn error of metabolism*", as Garrod used to designate some metabolic diseases, are usually recessive. In fact, there is some logic in these observations: the genes encoding metabolic enzymes are in general *haplosufficient* (50 % of normal levels are sufficient to complete the metabolic function), while the situation is radically different if the encoded polypeptide is involved in the differentiation of a specific tissue.

2.3.3 Epistasis and Pleiotropy

Many phenotypic traits are controlled by more than one gene, and, conversely, it is relatively common to observe that a given gene contributes to the phenotypic expression of one or several other genes. In the forthcoming chapters (in particular in Chap. 10, which is devoted to quantitative genetics) this point will be considered in detail. For the time being, we will just discuss a few examples that will help introducing two fundamental notions in genetics: *epistasis* and *pleiotropy*.

2.3.3.1 Epistasis

Epistasis characterizes a situation where the phenotypic expression of a gene (or allele) *A* depends on the presence, at some other loci in the genetic background (*B*, *C*, *D*), of one or several specific alleles that modify or suppress the classical

[31] A mutation that leads to the synthesis of a mutant protein that interferes or disrupts the activity of the wild-type protein in the multimer is called a *dominant-negative mutation*. A typical example is found in the syndrome of osteogenesis imperfecta (O.I. Type III) in which structurally defective type I collagen is formed.

phenotype of gene *A*. To put it in other words: epistasis is an interaction between non-allelic genes in which one gene suppresses or enhances the expression of another. The gene that is expressed is *epistatic* over the others genes, which are themselves *hypostatic*. Once more, the genes that are involved in the development of mouse coat color offer simple and didactic examples.

Exploiting the variety of alleles at the five major loci governing the mouse coat color (Agouti-*A*; Tyrosinase-*Tyr*; Brown-*Tyrp1*; Dilute-*Myo5a*; and Pink-eyed dilution-*Oca2*) one can generate a large collection of mice with a wide array of coat colors. However, sometimes it happens that the effects of a given mutant allele cannot be observed if another allele is present in the genome of the same animal. A mouse with a non-agouti, brown coat color (genotype *a/a*; $Tryp1^{b}/Tyrp1^{b}$) would appear "chocolate", unless the Tyr^{c} allele (which is at the *Tyr* locus on a different chromosome) is homozygous. In this latter case, the mouse would appear albino—i.e., completely white, and this is because the Tyr^{c} allele exhibits epistatic interaction with all other coat color genes. We know the reason: it is because the Tyr^{c} allele is non-functional. Thus, there is no tyrosinase synthesis, no melanin, and no pigment, be it black or yellow.

Another example of epistatic interaction in the mouse is between the $Mc1r^{e}$ allele (recessive yellow-Chr 8) at the locus encoding the melanocortin 1 receptor and the $Mlph^{ln}$ allele (leaden-Chr 1): mice with a $Mc1r^{e}/Mc1r^{e}$; $Mlph^{+}/-$ genotype have a deep yellow coat color, and mice with a $Mc1r^{+}/-$; $Mlph^{ln}/Mlph^{ln}$ genotype have a light gray coat color, like diluted, but these mice are indistinguishable from the $Mc1r^{e}/Mc1r^{e}$; $Mlph^{ln}/Mlph^{ln}$ mice. In short: $Mlph^{l}$ is epistatic to $Mc1r^{e}$ and the phenotypic effects of the recessive yellow allele are entirely suppressed by the phenotypic effects of leaden (Hauschka et al. 1968). Many mutations affecting enzymes of cellular metabolism exhibit epistatic effects, especially when they are in the same metabolic pathway.

Epistatic interactions are common with traits governing quantitative inheritance: the quantitative trait loci (QTLs). A heritable quantitative trait can be under the control of several genes with additive effects, the genes in question having different strengths in the determinism of the phenotype. When two alleles at different loci have a stronger effect on the phenotype than each allele individually, this is referred to as *synergistic epistasis*. The opposite situation also exists and is called *negative* or *antagonistic epistasis*.

When we described the epistatic effects of the Tyr^{c} (albino) allele on the expression of all other genes involved in coat color determinism, we assumed that this effect was the direct consequence of the expression (or non-expression, in this case) of the protein encoded by the gene (tyrosinase). The situation is sometimes much more subtle. For example, the mutant allele Apc^{Min} (at the adenomatous polyposis coli gene-Chr 18) is a dominant allele (although recessive lethal) that predisposes mice to the development of multiple intestinal neoplasia (Moser et al. 1990). However, some mouse strains are much more susceptible to this syndrome than others. Mice of the strain C57BL/6, for example, are severely affected and develop many intestinal tumors, while mice of the AKR strain, with the same

Apc^{Min} allele (congenic mice), develop only a few tumors. This dramatic phenotypic difference between the two inbred strains has been found to be the consequence of an epistatic interaction between the *Apc*^{Min} allele and another gene called *Modifier of Min* encoding a phospholipase A2 (*Pla2g2a*-Chr 4), itself with two alleles: *Pla2g2a*$^{Mom1\text{-}r}$ and *Pla2g2a*$^{Mom1\text{-}s}$. However, the *Pla2g2a* alleles have a phenotypic effect only when the *Apc*^{Min} allele is in the same genome. In other words, *Pla2g2a* is a modifier gene whose phenotypic expression is conditional to the presence of the *Apc*^{Min} allele. Such situations are very common in the mouse, and the *Apc*^{Min} allele has several other independent modifiers (Dietrich et al. 1993). The identification and study of modifier genes opens interesting avenues for unraveling the networks that determine robustness and resistance to certain diseases. Hence, we emphasize the importance of the use of pure inbred backgrounds in mouse models (see Chap. 9).

2.3.3.2 Pleiotropy

Pleiotropy describes a situation that is in fact extremely common in genetics: it simply means that a mutant allele has an effect on different phenotypic traits. In fact, if we carefully analyze most of the mutants with a deleterious phenotype, we would then discover that almost all of them in fact exhibit a panel of different phenotypes. The yellow allele (A^y) was identified because of its eye-catching phenotype, with a beautiful yellow coat color, but the mutant mouse exhibits many other phenotypes. For example, the mice are slightly diabetic, exhibit liver hypertrophy, and many become obese and sterile after the first few months of life. They are also more susceptible to several kinds of tumors than normal mice and are more aggressive.

If we consider that the products of most genes are involved in several cellular functions, pleiotropy seems to be more the rule than the exception. It simply means that the gene in question codes for a product that is used by various cells, or has a signaling function on various targets, or that the protein is an enzyme or a transcription factor that is involved in several pathways.

2.3.4 Penetrance and Expressivity

2.3.4.1 Penetrance

Penetrance is a term used to express the fraction of individuals of a given genotype that effectively exhibit the expected phenotype. Penetrance is usually expressed as a percentage. For example, if a particular dominant mutation has 80 % penetrance, then 80 % of the mice carrying the mutant allele will develop the phenotype and 20 % will look normal (Fig. 2.1).

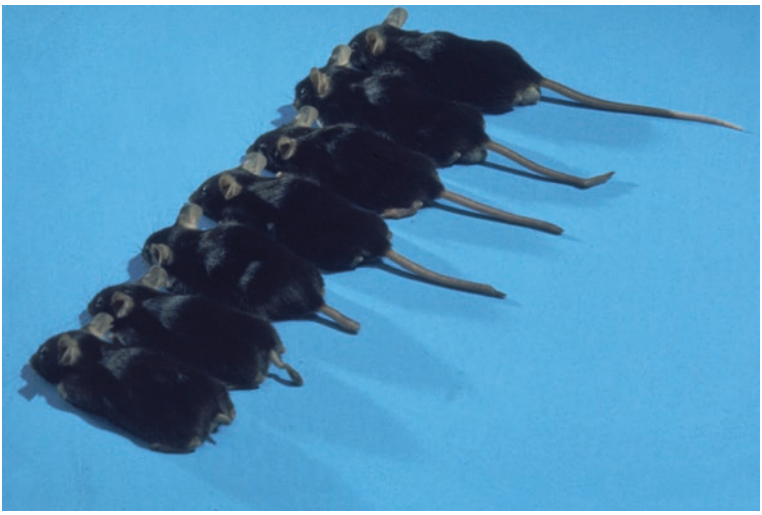


Fig. 2.1 *Penetrance and expressivity*. The figure illustrates the concepts of penetrance and expressivity. In this example, the mutation brachyury (*T*-Chr 17), affecting six out of the seven mice, exhibits great variations in expressivity; some mice have a tail longer than others, even if they all are clearly short-tailed. When a mouse with a short tail (genotype certainly *T*/+) is crossed with a normal mouse (+/+), the proportion of affected offspring is often lower than 50 %. Some mice have an extremely severe reduction of the tail, exhibit a *spina bifida*, and die at birth while others have an almost normal tail (normal overlaps). The penetrance characterizes the fraction of individuals of a given genotype that actually exhibit the phenotype typical of the mutant allele, irrespective of the degree of its expression. The expressivity characterizes the phenotypic variation among individuals having the same genotype. It is now well established that modifier genes influence the phenotypic expression of many mutant alleles. However, the action of these modifiers cannot explain all types of variations since phenotypic variations are also observed in inbred strains—as in the case illustrated here, where all the mice are from the same inbred strain. Variations in penetrance and expressivity are common for skeletal and eye mutations in all species

2.3.4.2 Expressivity

A genotype exhibits variable *expressivity* when individuals with that genotype differ in the extent to which they express the phenotype normally associated with that genotype. The best example illustrating the concept of expressivity and differentiate it from the concept of penetrance (which is not always easy) was provided by Danforth regarding a population of cats in Key West Island (a population also known as "Hemingway's cats"), in which a dominant mutation resulting in polydactyly is highly prevalent. Observing the cats in question, Danforth wrote, "*the polydactyly phenotype shows good penetrance, but variable expression*". This simply meant that a high percentage of cats indeed had extra toes, but the number and size of the extra toes varied from one animal to the next (Danforth 1947). Another example is the case of spotting in cattle. Observing a herd of cows of the Frisonne breed one may notice that, although all the cows are spotted (penetrance

is 100 %), the ratio black/white is highly variable from one cow to the next. The spotting is highly variable in shape (no surprise) but also in extension (which is more surprising). These are variations in expressivity of the spotting allele.

Although the examples we selected were from cats or cows, similar situations can be easily found in mice where mutations yielding extra digits and white spotting are common. In short, variable expressivity means that there is a large amount of phenotypic variation among individuals with the same genotype (Miko 2008).

The causes of penetrance and expressivity are not well understood. In the mouse, as well as in the rat, one can study the phenotypic expression of the same mutation in different genetic backgrounds and note more or less consistent differences, indicating the existence of a genetic component (modifier genes). However, in the same species, one can also observe phenotypic variations in animals having exactly the same mutation in exactly the same genetic background—meaning that genetic factors are not the only factors involved in the variation of penetrance or expressivity. In these conditions, it makes sense to consider that epigenetic factors or stochastic events are probably also at work. In Chap. 6, dealing with the epigenetic control of genome expression, we will discuss a situation where the coat color of mutant mice is strongly influenced by environmental factors.

Having control of the factors that determine expressivity is of major importance in human medicine, because many diseases with a genetic determinism (for example, cancers, neurological diseases, and skeletal abnormalities) often exhibit great variations in expressivity (Nadeau 2003).

2.4 Phenotyping Laboratory Mice: The Mouse Clinics

As we will explain in the chapters to come, researchers now have all the means and tools to create a great variety of alterations in the mouse genome; for example, they can switch off any gene they wish, and at any time. They can interfere (transitorily or not) with gene regulation, they can make all sorts of genetically engineered mice with cloned DNAs of their choice, etc. Of course, all of these alterations induced at the genome level are expected to result in changes at the phenotypic level in genetically modified animals, and the careful analysis of these phenotypic changes is obviously fundamental for the process of *genome annotation.*[32] However, the problem is that, though it is relatively easy to localize and precisely characterize a DNA sequence, especially nowadays, it is much more difficult to unambiguously establish the link between a DNA alteration and an abnormal phenotype. Examples are numerous where the knockout allele of a theoretically important gene was initially reported as having "no detectable phenotype," and this was to the great surprise (and sometimes to the disappointment) of its creator (Colucci-Guyon et al. 1994).

[32] Genome annotation consists of attaching biological information to a particular DNA sequence, or of establishing a link between a gene (or a small chromosomal region) and a given phenotype by any possible means.

Phenotyping has become one of the main concerns of mouse geneticists over the last 10 years and, mainly for this reason, many laboratories and institutions have developed what is now called a "*Mouse Clinic*". In these clinics, mouse mutants or strains are thoroughly analyzed for the greatest possible number of parameters using a panel of highly standardized phenotyping protocols. In most cases the basic protocols are focused on behavior, bone and cartilage development, neurology, clinical chemistry, eye development, immunology, allergy, steroid metabolism, energy metabolism, lung function, vision and hearing, pain perception, molecular phenotyping, cardiovascular analyses, and gross pathology. For example, the International Mouse Phenotyping Resource of Standardised Screens (IMPReSS) contains standardized phenotyping protocols, which are essential for the characterization of mouse phenotypes (see https://www.mousephenotype.org/impress). The use of standard procedures and defined protocols allows data to be comparable and shareable, and even allows interspecies comparisons to be performed, which may help in the identification of mouse models of human diseases.

References

Bradley A, Evans M, Kaufman MH, Robertson E (1984) Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. Nature 309:255–256

Brinster RL (1974) The effect of cells transferred into the mouse blastocyst on subsequent development. J Exp Med 140:1049–1056

Bruce HM (1959) An exteroceptive block to pregnancy in the mouse. Nature 184:105

Byers SL, Wiles MV, Dunn SL, Taft RA (2012) Mouse estrous cycle identification tool and images. PLoS ONE 7:e35538

Colucci-Guyon E, Portier MM, Dunia I, Paulin D, Pournin S, Babinet C (1994) Mice lacking vimentin develop and reproduce without an obvious phenotype. Cell 79:679–694

Condamine H, Custer RP, Mintz B (1971) Pure-strain and genetically mosaic liver tumors histochemically identified with the -glucuronidase marker in allophenic mice. Proc Natl Acad Sci USA 68:2032–2036

Danforth CH (1947) Heredity of polydactyly in the cat. J Heredity 38:107–112

De Repentigny Y, Kothary R (1996) An improved method for artificial insemination of mice–oviduct transfer of spermatozoa. Trends Genet 12:44–45

de Vries H (1910) Intracellular pangenesis (trans from German: Stuart Gager C). The Open Court Publishing Co., Chicago

Detlefsen JA (1921) A new mutation in the house mouse. Amer Nat 55:469–476

Dietrich WF, Lander ES, Smith JS, Moser AR, Gould KA, Luongo C, Borenstein N, Dove W (1993) Genetic identification of *Mom-1*, a major modifier locus affecting *Min*-induced intestinal neoplasia in the mouse. Cell 75:631–639

Eicher EM, Hoppe PC (1973) Use of chimeras to transmit lethal genes in the mouse and to demonstrate allelism of the two X-linked male lethal genes jp and msd. J Exp Zool 183:181–184

Fox JG, Barthold SW, Davisson MT, Newcomer CE, Quimby FW, Smith AL (2007) The mouse in biomedical research, 2nd edn. Elsevier, New York

Gardner RL, Lyon MF (1971) X chromosome inactivation studied by injection of a single cell into the mouse blastocyst. Nature 231:385–386

Glenister PH, Rall WF (2000) Cryopreservation and rederivation of embryos and gametes. In: Jackson IJ, Abott CM (eds) Mouse genetics and transgenics: a practical approach. Oxford University Press, Oxford

Grüneberg H (1952) The genetics of the mouse. Martinus Nijhoff, The Hague

Guénet JL, Babinet C (1978) The hairy ear mutation (*Eh*) is not cell lethal. Mouse News Letter 58:67
Guénet JL, Marchal G, Milon G, Tambourin P, Wendling F (1979) Fertile dominant spotting in the house mouse. J Hered 70:9–12
Hauschka TS, Jacobs BB, Holdridge BA (1968) Recessive yellow and its interaction with belted in the mouse. J Hered 59:339–341
Hedrich HJ (2012) The laboratory mouse, 2nd edn. Elsevier Academic Press, Amsterdam
Hogan ME, King LE Jr, Sundberg JP (1995) Defects of pelage hairs in 20 mouse mutations. J Investig Dermatol 104(5 Suppl):31S–32S
Illmensee K, Kaskar K, Zavos PM (2005) Efficient blastomere biopsy for mouse embryo splitting for future applications in human assisted reproduction. Reprod Biomed Online 11:716–725
Kamimura S, Inoue K, Ogonuki N, Hirose M, Oikawa M, Yo M, Ohara O, Miyoshi H, Ogura A (2013) Mouse cloning using a drop of peripheral blood. Biol Reprod 89:24
Kang L, Wang J, Zhang Y, Kou Z, Gao S (2009) iPS cells can support full-term development of tetraploid blastocyst-complemented embryos. Cell Stem Cell 5:135–138
Kaufman MH, O'Shea KS (1978) Induction of monozygotic twinning in the mouse. Nature 276:707–708
Kobayashi T, Yamaguchi T, Hamanaka S, Kato-Itoh M, Yamazaki Y, Ibata M, Sato H, Lee YS, Usui J, Knisely AS, Hirabayashi M, Nakauchi H (2010) Generation of rat pancreas in mouse by interspecific blastocyst injection of pluripotent stem cells. Cell 142:787–799
Leckie PA, Watson JG, Chaykin S (1973) An improved method for the artificial insemination of the mouse (*Mus musculus*). Biol Reprod 9:420–425
Luo C, Zuñiga J, Edison E, Palla S, Dong W, Parker-Thornburg J (2011) Superovulation strategies for 6 commonly used mouse strains. J Am Assoc Lab Anim Sci 50:471–478
McLaren A, Molland P, Signer E (1995) Does monozygotic twinning occur in mice? Genet Res 66:195–202
Miko I (2008) Phenotype variability: penetrance and expressivity. Nature Education 1:137
Mintz B (1962) Formation of genetically mosaic mouse embryos. Am Zool 2:432
Mintz B, Baker WW (1967) Normal mammalian muscle differentiation and gene control of isocitrate dehydrogenase synthesis. Proc Natl Acad Sci USA 58:592–598
Mintz B, Illmensee K (1975) Normal genetically mosaic mice produced from malignant teratocarcinoma cells. Proc Natl Acad Sci USA 72:3489–3585
Mintz B, Silvers W (1967) "Intrinsic" immunological tolerance in allophenic mice. Science 158:1484–1487
Mochida K, Hasegawa A, Taguma K, Yoshiki A, Ogura A (2011) Cryopreservation of Mouse Embryos by Ethylene Glycol-Based Vitrification. J Vis Exp 57:e3155. doi:10.3791/3155
Moser AR, Pitot HC, Dove WF (1990) A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. Science 247:322–324
Nadeau JH (2003) Modifier genes and protective alleles in humans and mice. Curr Opin Genet Dev 13:290–295
Nagy A, Gertsenstein M, Vintersten K, Behringer R (2003) Manipulating the mouse embryo, a laboratory manual, 3rd edn. Cold Spring Harbor Press, New York
Nakagata N (2011) Cryopreservation of mouse spermatozoa and in vitro fertilization. Methods Mol Biol 693:57–73
Nakagata N, Takeo T, Fukumoto K, Haruguchi Y, Kondo T, Takeshita Y, Nakamuta Y, Umeno T, Tsuchiyama S (2014) Rescue In Vitro Fertilization Method for Legacy Stock of Frozen Mouse Sperm. J Reprod Dev 60(2):168–171
Nishizono H, Shioda M, Takeo T, Irie T, Nakagata N (2004) Decrease of fertilizing ability of mouse spermatozoa after freezing and thawing is related to cellular injury. Biol Reprod 71:973–978
Ogonuki N, Inoue K, Hirose M, Miura I, Mochida K, Sato T, Mise N, Mekada K, Yoshiki A, Abe K, Kurihara H, Wakana S, Ogura A (2009) A high-speed congenic strategy using first-wave male germ cells. PLoS ONE 4:e4943. doi:10.1371/journal.pone.0004943
Ogonuki N, Inoue K, Ogura A (2011) Birth of normal mice following round spermatid injection without artificial oocyte activation. J Reprod Dev 57:534–538

Ogonuki N, Mochida K, Miki H, Inoue K, Fray M, Iwaki T, Moriwaki K, Obata Y, Morozumi K, Yanagimachi R, Ogura A (2006) Spermatozoa and spermatids retrieved from frozen reproductive organs or frozen whole bodies of male mice can produce normal offspring. Proc Natl Acad Sci USA 103:13098–13103

Ogura A, Ogonuki N, Inoue K, Mochida K (2003) New microinsemination techniques for laboratory animals. Theriogenology 59:87–94

Ogura A, Ogonuki N, Miki H, Inoue K (2005) Microinsemination and Nuclear Transfer Using Male Germ Cells. Int Rev Cytol 246:189–229

Olds-Clarke P (1989) Sperm from tw^{32}/+ mice: capacitation is normal, but hyperactivation is premature and nonhyperactivated sperm are slow. Dev Biol 131:475–482

Panthier JJ, Guénet JL, Condamine H, Jacob J (1990) Evidence for mitotic recombination in W(ei)/+ heterozygous mice. Genetics 125:175–182

Papaioannou VE, McBurney MW, Gardner RL, Evans MJ (1975) Fate of teratocarcinoma cells injected into early mouse embryos. Nature 258:70–73

Perez CJ, Jaubert J, Guénet J-L, Barnhart KF, Ross-Inta CM, Quintanilla VC, Aubin I, Brandon J, Otto N, DiGiovanni J, Gimenez-Conti I, Giulivi C, Kusewitt DF, Conti CJ, Benavides F (2010) Two hypomorphic alleles of mouse *Ass1* as a new animal model of citrullinemia type I, and other hyperammonemic syndromes. Am J Pathol 177:1958–1968

Silvers WK (1979) The coat colors of mice: a model for mammalian gene action and interaction. Springer, Berlin

Stewart CL, Gadi I, Bhatt H (1994) Stem cells from primordial germ cells can reenter the germ line. Dev Biol 161:626–628

Sztein J, Vasudevan K, Raber J (2010) Refinements in the cryopreservation of mouse ovaries. J Am Assoc Lab Anim Sci 49:420–422

Sztein JM, Farley JS, Mobraaten LE (2000) In vitro fertilization with cryopreserved inbred mouse sperm. Biol Reprod 63:1774–1780

Taft RA, Low BE, Byers SL, Murray SA, Kutny P, Wiles MV (2013) The perfect host: a mouse host embryo facilitating more efficient germ line transmission of genetically modified embryonic stem cells. PLoS ONE 8:e67826. doi:10.1371/journal.pone.0067826

Tarkowski A (1998) Mouse chimaeras revisited: recollections and reflections. Int J Dev Biol 42:903–908

Tarkowski AK (1961) Mouse chimaeras developed from fused eggs. Nature 190:857–860

Theiler K (1972) The house mouse. Springer, New York

Van der Lee S, Boot LM (1956) Spontaneous pseudopregnancy in mice II. Acta Physiol Pharmacol Neerl 5:213–215

Vergara GJ, Irwin MH, Moffatt RJ, Pinkert CA (1997) In vitro fertilization in mice: Strain differences in response to superovulation protocols and effect of cumulus cell removal. Theriogenology 47:1245–1252

Wakayama S, Kohda T, Obokata H, Tokoro M, Li C, Terashita Y, Mizutani E, Nguyen VT, Kishigami S, Ishino F, Wakayama T (2013) Successful serial recloning in the mouse over multiple generations. Cell Stem Cell 12:293–297

Wakayama T, Perry AC, Zuccotti M, Johnson KR, Yanagimachi R (1998) Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. Nature 394:369–374

West JD, Frels WI, Papaioannou VE, Karr JP, Chapman VM (1977) Development of interspecific hybrids of Mus. J Embryol Exp Morphol 41:233–243

Whitten WK (1956) Modification of the oestrous cycle of the mouse by external stimuli associated with the male. J Endocrinol 13:399–404

Whittingham DG (1968) Fertilization of mouse eggs in vitro. Nature 220:592–593

Whittingham DG, Leibo SP, Mazur P (1972) Survival of mouse embryos frozen to −196 degrees and −269 degrees C. Science 178:411–414

Wilmut I (1972) The effect of cooling rate, warming rate, cryoprotective agent and stage of development on survival of mouse embryos during freezing and thawing. Life Sci II 11:1071–1079

Wolfe HG (1967) Artificial insemination of the laboratory mouse (Mus musculus). Lab Anim Care 17:426–432

Didactic movie

http://www.jove.com/video/3155/cryopreservation-mouse-embryos-ethylene-glycol-based

Reference Book

Papaioannou VE, Behringer R (2005) Mouse phenotypes: a handbook of mutation analysis. CSHL Press, New York, p 235

Chapter 3
Cytogenetics

3.1 Introduction

Cytogenetics, as the name indicates, lies at the intersection between cell biology and genetics. It came into being as an independent discipline after the advent of the chromosomal theory of heredity at the beginning of the twentieth century, when W.S. Sutton and T.H. Boveri (then T.H. Morgan) identified the chromosomes as the physical structures on which the genes were anchored (1902–1915).[1]

Cytogenetics deals with all aspects of chromosomes biology: their morphology and structure, their number, and their behavior during mitosis and meiosis. It also deals with the pathology and functional changes associated with accidental variations in chromosome number or structure.

For a long time cytogenetics remained a rather descriptive discipline, but in recent years, particularly after the development of highly sophisticated staining techniques and because of the availability of enormous collections of chromosomal rearrangements, it has contributed to the development of genetic maps, to a better understanding of the developmental consequences of chromosomal aberrations in humans (Down syndrome, for example), and to the elucidation of some fundamental biological questions such as genomic imprinting (see Chap. 6).

With the recent advent of fluorescence in situ hybridization (FISH) techniques, cytogenetics has changed profoundly in the sense that it is now possible to visualize the chromosomes, or a specific region of them, in the interphasic nucleus. Taking advantage of the possibilities offered by such "interphasic cytogenetics", it has thus become possible to obtain information about chromosome number and structure in all tissues, at any time, independent of the cell cycle. In turn, interphasic cytogenetics has made possible the development of "functional cytogenetics", providing information relative to some epigenetic mechanisms of gene

[1] Thomas H. Morgan was awarded the Nobel Prize in Physiology or Medicine in 1933 for his discoveries concerning the role played by the chromosomes in heredity. Morgan proposed that each chromosome contains a collection of small units called "genes" that were linearly arranged on the chromosomes.

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_3

regulation; for example, in relation to imprinting. Finally, over the past few years, cytogenetics has benefited from the exuberant development of the molecular techniques of genetic engineering in embryonic stem cells (ES cells) (see Chap. 8). As a result, virtually any chromosomal rearrangement (deletion, duplication, transposition, inversion, etc.) of interest to researchers can now be designed in silico, engineered in vitro, and ideally analyzed in the context of a living mouse.

3.2 The Chromosomes of the Mouse

A simple way to observe mouse chromosomes is to collect a bone marrow sample and disperse it into tissue culture medium, to which is added 1–2 drops of a 0.025 % colchicine solution. After 20–30 min of such treatment, all actively dividing cells of the bone marrow are blocked at the metaphase stage because the drug prevents the synthesis of spindle fibers and, therefore, stops mitosis in metaphase. If these arrested cells are transferred into a hypotonic solution (for example, 0.56 % KCl) for a few minutes and then placed on a glass slide, they swell and burst open, and the chromosomes appear spread out on the slide. They can then be fixed with a simple fixative medium (for example, a 3:1 cold mixture of methanol and glacial acetic acid) and observed under the microscope, either in phase contrast or after staining, and counted. Such a simple technique allowed it to be shown, almost a century ago, that the normal laboratory mouse has 40 chromosomes (Cox 1926).[2] The word "chromosome" comes from the Greek *chroma* meaning color and *soma* meaning body, referring to their tendency to be strongly stained by stains like orcein or Sudan black (Fig. 3.1).

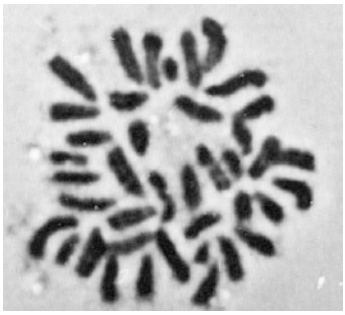


Fig. 3.1 *Spermatogonial metaphase of the mouse* (Sudan black squash). In this figure, which was published in the book *Biology of the Laboratory Mouse* (Dover Publications, 1966, Chap. 7), the 40 chromosomes of the mouse are clearly visible but it is very difficult to specifically identify the elements of the different pairs

[2] A technique for rapid chromosome staining is provided at: http://www.jax.org/cyto/marrow_preps_alt.html.

For many years, chromosomes were represented as a core of proteins, mostly but not only histones, to which a filament of supercoiled DNA was tightly attached, giving each chromosome its specific shape. Nowadays, this folding of DNA into coils and loops of increasing diameter is being reconsidered and scientists have several alternative descriptions of the intimate organization of the DNA molecule inside the mitotic chromosomes (Uhlmann 2013). Histones serve as a frame to hold the association in a compact structure and also to control DNA transcription, as we will discuss later. This association of DNA with histones, and some other proteins of the polycomb group, is given the generic name of *chromatin.* Two types of chromatin can be distinguished at the cellular level: (i) *euchromatin*, whose structure is open and moderately condensed, allowing the transcription of a variety of RNAs, and (ii) *heterochromatin*, whose structure is highly condensed, impeding DNA transcription. *Heterochromatin* contains repetitive sequences and a special form of it is *constitutive heterochromatin*, which is located mostly at or near the centromeres. Cell biologists hypothesize that constitutive heterochromatin probably has only a structural function, while euchromatin has a function in gene expression and regulation.

When cells enter the process of mitotic division the chromatin becomes progressively condensed, the chromosomes become shorter in size and clearly visible. They then split into two *chromatids* that remain attached for a period of time by their centromere. The chromatids bind to the microtubule of the mitotic spindle by a special protein structure: the *kinetochore*. Then, the centromeres themselves split, the mitotic process goes on, and the chromatids become individualized in chromosomes that are pulled apart to the opposite poles of the cell. Finally, this results in two daughter cells, each with its own replica of the original set of chromosomes. After mitotic division, the cells return to interphase, the chromatids uncoil, and DNA can again be transcribed.

The chromosomes vary in size according to the stage of the cell cycle in which they are observed and the degree of chromatin condensation. They also vary in shape and exist either as single linear strands (unduplicated chromosomes) or as duplicated chromosomes, just before anaphase, when the two chromatids are still joined by their centromere.

As long as it is not duplicated, each chromosome contains a single and unique molecule of supercoiled DNA. The mammalian genome is thus fragmented into discrete elements that consist of a linear array of genes encoding RNAs and proteins, interspersed with noncoding DNA. They also contain a special DNA segment that constitutes the *centromere* and two special stretches at their ends: the *telomeres*.

In the mouse, the centromeres (symbol *Cen*) consist of a relatively large array of repetitive heterochromatin containing satellite DNA (see Chap. 5), where the sequence within the individual repeats is similar but not identical. The normal histone H3 is replaced by a variant (CENP-A) that is believed to be important for the assembly of the kinetochore.

The telomeres (symbol *Tel*) consist of a variable number of the tandemly repeated motif–5′-TTAGGG-3′–bound to specialized proteins and measuring

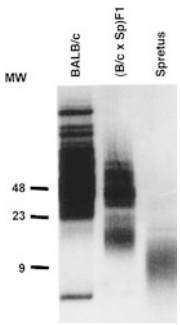


Fig. 3.2 *Telomere length*. Telomere length of BALB/c (*M. musculus*), *M. spretus*, and (BALB/c × *M. spretus*) F1 somatic cells. Genomic DNA was subjected to restriction digestion with the enzymes *Hinf*I and *Rsa*I, and analyzed by pulse field electrophoresis. BALB/c mice have "long" telomeres, while *M. spretus* mice have "short" telomeres (from Zhu et al. 1998)

up to 40 kb in the laboratory mouse. For many years, the telomeres had been considered as mere insulators at the end of the chromosomal DNA filament, playing a role similar to the role played by the plastic caps that are molded at the ends of our shoelaces; to prevent their ends from being damaged or accidentally tied to each other by the DNA repair enzymes. Nowadays, some geneticists hypothesize that the enzyme telomerase, whose function is, in particular, to control the length of telomeres by adding new–5′-TTAGGG-3′–monomers, plays a crucial role in the control of cell senescence, and that telomere length is a way of monitoring the cell's replicative potentialities, with long telomeres being indicative of greater potentialities (Blasco 2005; Sahin and De Pinho 2010). However, this point is still debated because the comparison of telomere sizes in mice from different species but of the same genus *Mus* does not support the hypothesis that telomere shortening is correlated with senescence in mice (Kim et al. 2003). Laboratory mice (*Mus musculus*), for example, have long telomeres, while wild mice of the species *Mus spretus* have short ones (i.e., like human) but a very similar lifespan and behavior (Zhu et al. 1998). Similarly, mice homozygous for a knockout allele of *Tert* (the gene encoding telomerase reverse transcriptase) do not exhibit any phenotypes related to accelerated senescence (Fig. 3.2).

3.3 Identifying the Chromosome Pairs: The Normal Karyotype

In most mammalian species, including human, one can generally observe three kinds of chromosomes, depending on the position of the centromere. When the centromere is roughly in the middle of the chromosome, the latter is said to be *metacentric*. When it is slightly shifted and divides the chromosome unequally,

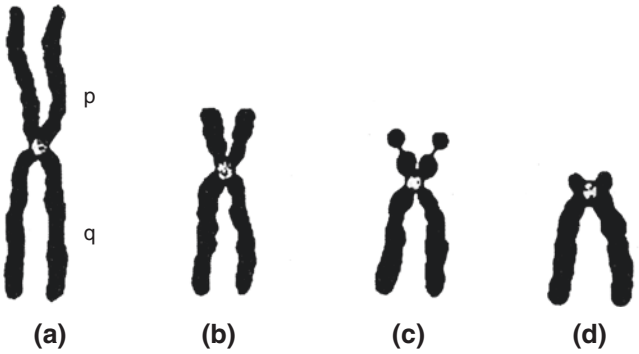


Fig. 3.3 *Sorting out the different mouse chromosomes*. Chromosomes are generally classified according to their size and shape. **a** Metacentric; **b** Sub-metacentric; **c** Acrocentric; **d** Telocentric. Telocentrics are an extreme form of acrocentrics. The centromeric index is computed based on the ratio $p/p + q$, where p is the size of the short arm and q the size of the long arm. Metacentric chromosomes have a centromeric index of 0.5 ($p/p + q = 0.5$). Acrocentric chromosomes have a centromeric index <0.5. Traditional laboratory mice have acrocentric chromosomes only

with long arms and short arms, the chromosome is said to be *sub-metacentric*. Finally, when the centromere is subterminal, the chromosome is said to be *acrocentric*. In literature from the 1960s some chromosomes were depicted as *telocentric* with their centromere completely shifted to one end. Specialists now consider that telocentric chromosomes are unstable structures and probably do not exist in reality. Chromosomes are also classified according to two criteria: first, their global size, and second, their centromeric index. The centromeric index is computed based on the ratio $p/p + q$, where p (from the French *petit* meaning small) is the size of the short arm and q the size of the long arm (from the French *queue* meaning tail). Metacentric chromosomes have a centromeric index of 0.5 ($p/p + q = 0.5$), while acrocentrics have a centromeric index $\ll 0.5$ (Fig. 3.3).

The chromosome set of a given species is generally presented with the chromosomes being displayed according to their size, from the largest to the smallest, and, when of the same size, according to the centromeric index. Once arranged as described, the chromosome set of a given species represents its *karyotype*. The karyotype is a fundamental parameter that is generally unique to a species. Thus, the karyotype of the normal laboratory mouse (MMU or sometimes Mmu for *M. musculus*) consists of 40 acrocentric chromosomes, i.e., 19 autosomes plus an X and a Y (in short 40,XY). Unfortunately, unless these chromosomes are stained with a special technique (see below), it is difficult to individually identify the members of the different pairs. Chromosome 1 represents 7.5 % of the mouse haploid genome, the X chromosome is the fifth in size and represents 6.60 % of the haploid genome, and chromosome 19 is the smallest entity and represents approximately 2.3 % of the haploid genome. The Y chromosome has the same size as chromosome 18 (3.5 %) (Fig. 3.4).

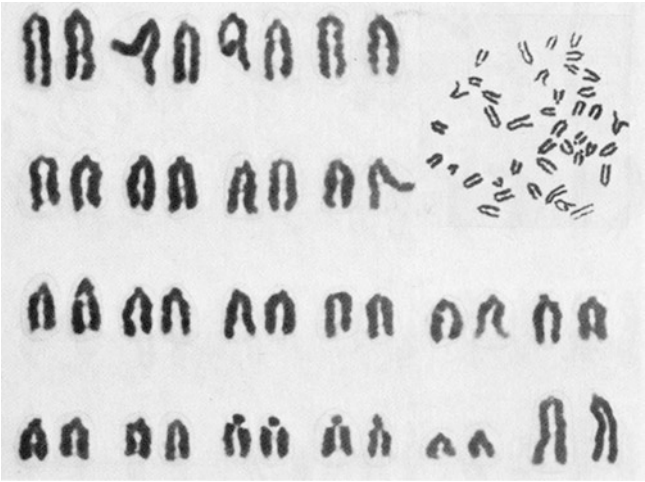


Fig. 3.4 *Ideogram of the mouse chromosomes*. Ideogram of the metaphasic chromosomes of a female mouse (from the book *Biology of the Laboratory Mouse*, Dover Publications, 1966). The two chromosome 19s are clearly identifiable by their short size on this preparation

Wild mice sometimes exhibit variations in their karyotype concerning the total number of centromeres but not the total number of arms.[3] For example, some wild mice living in Western Europe (Switzerland, southern Germany, northern Italy) and North Africa have a karyotype with a variety of metacentric chromosomes. All these variations are considered normal and do not affect the fertility and/or viability of the animals as long as they are homozygous. However, when these mice are crossed with laboratory strains, whose karyotypes are composed of acrocentric chromosomes only, they produce normal, healthy F1, but the latter are almost always completely sterile because they produce gametes with abnormal (unbalanced) sets of chromosomes ($\neq$n). We will come back to this point later in this chapter.

During the early 1970s, considerable progress was made concerning the techniques used to stain the chromosomes. The first of these techniques was reported for human chromosomes by T. Caspersson and colleagues. It makes use of the fluorescent dye quinacrine (Caspersson et al. 1970, 1972; Miller et al. 1971; Miller and Miller 1975) and yields a series of lightly and darkly stained bands on the chromosome arms. This banding pattern is characteristic of each and every pair of chromosomes, or nearly so. These bands are called the Q bands and the technique is still in use, although infrequently.

Another popular technique was developed almost simultaneously (Sumner et al. 1971), and was based on the controlled enzymatic digestion of chromatin with either trypsin or chymotrypsin, followed by conventional staining with the Giemsa dye. The banding pattern characteristic of this technique rapidly became popular

[3] The total number of chromosome arms per set of chromosomes is called the *fundamental number* (*nombre fondamental*) after Matthey.

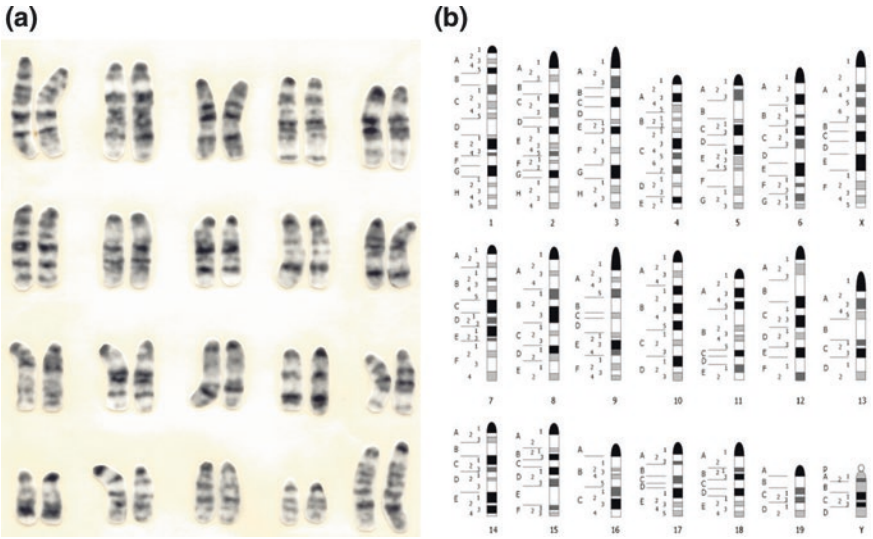


Fig. 3.5 **a** *The mouse karyotype*. The mouse karyotype stained by the Giemsa dye after trypsin digestion. This G-banding pattern allows the individual identification of each chromosome pair with only a few ambiguities (8–12; 9–13). (This figure is courtesy of Dr. Heinz Winking, Insitut für Biologie, Medizinische Universität zu Lübeck, Germany). **b** *The mouse ideogram* is a schematic and standardized representation of the different chromosome bands (Figure from Dr. David Adler, University of Washington Department of Pathology). http://www.pathology.washington.edu/research/cytopages/idiograms/mouse

and is referred to as G-banding: dark regions are heterochromatic and AT-rich, while light regions are euchromatic and GC-rich (Sawyer et al. 1987). This technique using trypsin/Giemsa is still used today. The G-banding pattern is highly reproducible and the schematic representation of the bands in the mouse is called the *ideogram*. The Q-banding pattern is very similar to the G pattern, and, being technically more difficult to achieve, this explains why it has been progressively replaced by the G-banding technique.

These methods will normally produce around 350 bands in a normal mouse karyotype (Nesbitt and Francke 1973) (Fig. 3.5).

Using the same Giemsa staining with some technical variations, other banding patterns have been described in the past:

- R-banding, which is the reverse of G-banding (the R stands for "reverse"). The dark regions are euchromatic (GC-rich regions), while the bright regions are heterochromatic (AT-rich regions).
- C-banding: Giemsa binds to constitutive heterochromatin and stains the centromeres.
- T-banding that visualizes the telomeres.

All these techniques, except for G-banding, have now been replaced by more modern methods that have been transferred from human cytogenetics. These

techniques make use of different fluorochromes and a variety of (human or mouse) chromosome-specific DNA probes and they are, at the same time, more reliable and easier to standardize. They can stain specifically an entire mouse chromosome, if the DNA probe is specific for this mouse chromosome, or the fragment of a mouse chromosome that is homologous to a human chromosome, if the DNA probe is specific for a human chromosome. These techniques for "chromosome painting" are now standardized and are very helpful for the transposition of information from one species to the next, and can be applied to any domestic species if necessary. Chromosome painting is also useful for analyzing chromosome rearrangements.

Fluorescence in situ hybridization (FISH) is an interesting technique when used for the localization of a specific gene or transgene in the mouse. One can, for example, stain the karyotype with either the G-banding technique or any other chromosome painting technique, and look for the localization of a specific DNA sequence (a gene or a transgene, for example) with another probe labeled either with a fluorochrome or a radioactive isotope. The localization of a transgene, which has always been problematic, has been greatly simplified by the use of such staining techniques even though, in many cases, the localization is only roughly assessed (Fig. 3.6).

Spectral karyotyping is a molecular technique used to visualize the different pairs of chromosomes of a given species in different colors (Liyanage et al. 1996). This is achieved by labeling chromosome-specific DNA with different fluorophores and then using these probes for hybridization with the chromosomes.

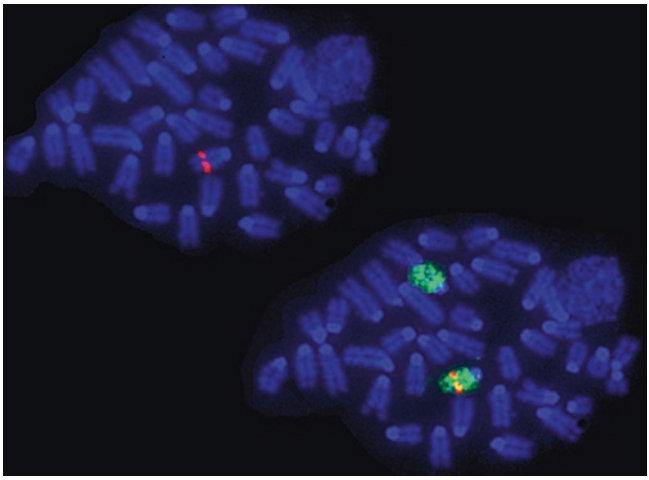


Fig. 3.6 *Mapping a transgenic insertion*. Observation of a transgene in the mouse by fluorescence in situ hybridization with a molecular probe (stained in *red* by rhodamine). The transgenic insertion is on mouse chromosome 12 that appears labeled *green* by chromosome-specific painting. The technique is of great help for the derivation of a mouse strain homozygous for the transgenic insertion. (This figure is courtesy of Dr. M.G. Mattei, Hôpital de la Timone, Marseille)

Fig. 3.7 *Spectral karyotyping*. The figure represents the karyotype from a normal male mouse (strain C57BL/6) generated by using the multi-color FISH cytogenetic method called spectral karyotyping or SKY. The colors to the left of each black and white (G-banded) chromosome (derived from inverted-DAPI staining) are the RGB (red–green–blue) display of the fluorochromes. The pseudo-color to the right of the Giemsa-banded chromosomes are referred to as the classified colors and are derived from a mathematical algorithm that translates the wavelengths of each chromosome, for each pixel, and converts the wavelength chromosome-specific assignments into these classified colors. With this technique, cytogeneticists are now able to visualize complex karyotypes which involve multi-chromosomal rearrangements in an unambiguous manner. Contrary to traditional black-and-white karyotypes, visualization of chromosomal rearrangements with spectral karyotyping is straightforward, as one or more colors will show within a single chromosome. (This figure is courtesy of Drs. Hesed M. Padilla-Nash and Thomas Ried, Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA) (Liyanage et al. 1996; Green and Ried 2011)

Because there is a limited number of spectrally distinct fluorophores, a combinatorial labeling method is used to generate different colors. Spectral differences generated by combinatorial labeling are captured and analyzed by using an interferometer attached to a fluorescence microscope and then processing the pictures with a computer program that assigns a pseudo-color to each spectrally different combination, allowing the visualization of the individually colored chromosomes (Fig. 3.7).

Finally, taking advantage of the numerous variations in mouse karyotypes that are easily available nowadays, mouse chromosomes have been partially sorted by using flow cytometry. This has enabled the preparation of chromosome-specific DNA probes useful for mapping projects or for labeling (Baron et al. 1990).

3.4 Meiosis and Gametogenesis

Meiosis is a fundamental event in the life cycle of organisms reproducing sexually because, with the production of gametes, it marks the end of diplophase and the beginning of haplophase. Meiosis has been extensively studied and appears to

be remarkably similar across different species. In some aspects it resembles mitosis, another crucial step in the somatic cell cycle, but it has several fundamental differences.

1. While mitosis occurs in all types of cells and tissues, at least when the embryo differentiates and develops, meiosis occurs exclusively in the gonads, and more precisely in certain cells of the germinal lineage.
2. The end products of mitosis are two diploid (2n) daughter cells, which are absolutely identical from a genetic point of view, except in rare cases. These cells have chromosomes that are similar in number and structure to those in the mother cell and carry the same genes with the same linear ordering. In meiosis the situation is radically different: there are four daughter cells instead of two, and these cells are genetically unique in the sense that they have a unique, de novo assortment of the parental alleles and their chromosome number is halved (they are haploid).
3. During meiotic prophase (*diplonema—diakinesis*) the chromosomes are duplicated, creating two exact copies (or sister chromatids), but remain attached by their centromere. The maternal and paternal chromatids then pair and exchange segments by homologous recombination, leading to a patchwork between the maternal and paternal versions of the chromosome. The pairing between homologous chromosomes (between the two pairs of sister chromatids) is mediated by a protein structure known as the *synaptonemal complex*. This structure is temporary and disappears during late prophase. This inter-chromatid recombination (*chiasmata*) is specific to the meiotic process and is an efficient mechanism for the generation of diversity.

In male mice, meiosis occurs in spermatocytes I, producing spermatocytes II that later differentiate progressively into spermatids and finally become mature sperm cells (spermatozoa). Spermatogenesis in the mouse starts at about 3 weeks of age, puberty is reached by 6 weeks, and by 8 weeks of age a male is normally fully fertile. The duration of spermatogenesis is shorter in the mouse than in most other mammalian species and the time taken by spermatogonia to become mature spermatids, which are released into the lumen, is only 5 weeks. In theory, a single A1 spermatogonium gives rise to 256 sperm cells, but there is some attrition of cells during spermatogenesis and the actual number is smaller than this theoretical maximum (Fig. 3.8).

In female mice, meiosis begins immediately after the migration of the primordial germ cells into the ovary while the young female is still in utero. It is interrupted at the diplotene stage of meiosis I and the cells remain resting for a very long period. Then, stimulated by a burst of gonadotropin hormones released by the pituitary gland, groups of oocytes resume meiosis I and stop again at meiosis II until fertilization. At fertilization, when a spermatozoon penetrates the oocyte, meiosis is completed with the expulsion of the second polar body. In female mice, unlike in males, meiosis is a discontinuous process and its end-products are very unequal with two very small cells, a pronucleus I (first polar body with 2n chromosomes) and a pronucleus 2 (second polar body with n chromosomes), and one enormous cell, the oocyte with n chromosomes. However, it is important to note that, in the female, haploidy of the gamete is

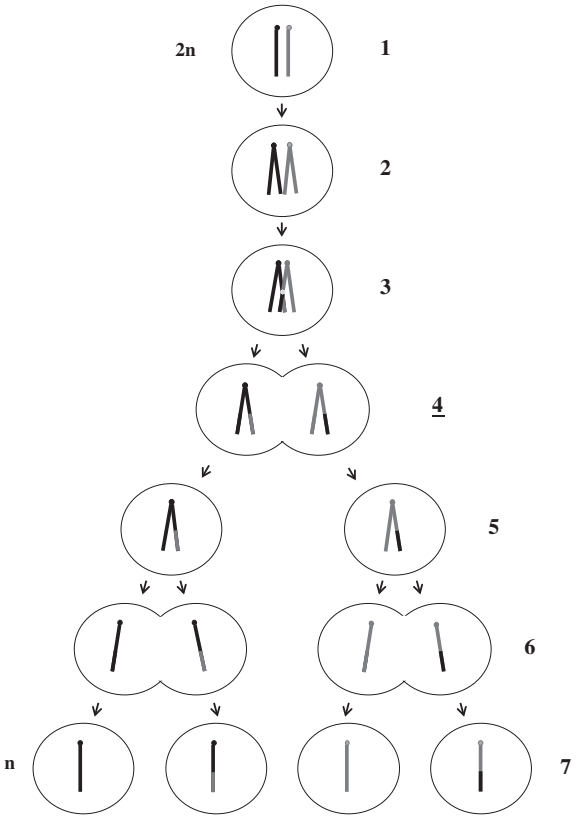


Fig. 3.8 *Schematic representation of the meiotic process*. *1*. The oogonia or spermatogonia stem cells are diploid (2n chromosomes). *2*. After one round of DNA synthesis, the chromosomes duplicate and create two exact copies of both the paternal and maternal chromosomes. These two copies, the chromatids, remain attached by their centromere. *3*. The maternal and paternal chromosomes, although divided in chromatids, pair and exchange parts by homologous recombination (crossing over). Chiasmas (*) can be observed at this stage. *4*. The centromere of each chromosome pair does not divide, but binds to the spindle fibers. They then split and the spindle fibers pull the chromatids to the opposite pole of the cell. This is the first disjunction. *5*. Sister chromatids remain together and form another equatorial plaque. *6*. The centromeres split and the individual chromosomes migrate to the opposite pole of the cell. This is the second disjunction. *7*. Four haploid cells are formed with only n chromosomes. Some of these chromosomes are recombinant and have genes from both parental chromosomes. In the male, the four products of meiosis are equivalent (spermatocytes II). In the female, one of the daughter cells at step 5 degenerates and forms the first polar body (with 2n chromosomes). Another one degenerates at step 7, when the spermatozoon penetrates the oocyte, and forms the second polar body (n chromosomes). The formation of the polar body is a random event and does not depend upon the genetic makeup of the cell (with a possible exception for XO females—see text)

only virtual because the expulsion of the second polar body (and its n chromosomes; late anaphase II) is triggered by the penetration of the spermatozoon into the oocyte.

Meiosis is a complex process involving several critical steps. Errors can occur at many of the steps, often leading to severe or even lethal abnormalities in the embryo. This will be the topic of the next sections in this chapter.

3.5 Variations in Chromosome Number

Following defective gametogenesis or abnormal fertilization, variations in chromosome number may occur accidentally and an embryo may then start its development with a set of chromosomes different from the fundamental diploid number 2n. These aberrations, called *heteroploidies*, are common to all mammalian species including human and are, in most instances, incompatible with normal development and often result in abortions. In the mouse, some heteroploidies represent interesting models for understanding the homologous human pathologies.

Geneticists sort these heteroploidies into two different subcategories according to the number of chromosomes involved. *Euploid heteroploidies* designate the case where the number of chromosomes in the conceptus is a multiple of the haploid number (n) (i.e., n, 3n, 4n, etc.). *Aneuploid heteroploidies* correspond to all other cases where there is a deviation from the normal 2n (for example, $2n + 1$, $2n - 1$, $2n + 1 + 1$ etc.).

3.5.1 The Euploid Heteroploidies

3.5.1.1 Haploidy (n)

In natural conditions, haploid embryos occur spontaneously but, as a rule, they are lethal at a very early stage of development. As we will discuss later (Chap. 6), a mammalian embryo cannot develop to term unless some of its chromosomes come from a male parent and others from a female parent. This, of course, represents a serious constraint on the development of haploid embryos.[4]

However, considering that haploid organisms have been created in several species, including vertebrates such as the medaka fish (*Oryzias latipes*), experiments have also been undertaken in the mouse. Haploid embryonic stem cells (ESCs) have been produced in vitro that were derived from parthenogenetic or androgenetic haploid embryos of several inbred mouse strains, collected at the blastocyst stage.

These haploid ESCs have been proven capable of a differentiation potential similar to that of diploid ESCs, and some have been used for genetic screening as well as for the production of homozygous mutant animals. These cells, however, are unstable and often spontaneously return to the diploid state. Two publications can be recommended concerning this subject (Leeb and Wutz 2011; Zhang and Teng 2013).

[4] UpD are exceptions that will be discussed later in this chapter and in Chap. 6.

3.5.1.2 Triploidy (3n)

Triploidy represents 2–3 % of human pregnancies and culminates in early spontaneous miscarriages. These heteroploidies result either from digyny (the basic 2n plus an extra haploid set of maternal origin), and originate through errors in meiosis II, or, in the majority of cases, from diandry (2n plus an extra haploid set of paternal origin) and originate from dispermy. Births of living triploid infants have been recorded, but these neonates suffer from multiple developmental defects and die shortly after birth.

In mice, micromanipulatory techniques have been used to produce digynic ($2n^M + n^P$) and diandric ($n^M + 2n^P$) triploid embryos (Niemierko 1981). These conceptuses revealed an ability to implant once transplanted into the uterus of suitable recipients and some developed up to the 15- to 25-somite stage. However, here again, aneuploid embryos appeared considerably smaller than euploids analyzed at the same stages of development. In the mouse, in contrast to what was described in human, digynic triploid conceptuses showed poorer embryonic development than the diandrics, but both were morphologically abnormal (Kaufman et al. 1989). Chimeric embryos created from triploid cells and normal diploid cells can progress to term, and some can even survive, depending on the ratio of the 3n/2n cells.

3.5.1.3 Tetraploidy (4n)

The generation of tetraploid embryos by electrofusion of 2n blastomeres of mouse embryos at the 2-cell or 4-cell stages of development has been reported by Kubiak and Tarkowski (1985). By applying a pulse of current, the authors succeeded in generating tetraploid embryos that could develop to the blastocyst stage but not further. Since these early experiments, tetraploid blastocysts have been used as recipients for the production of mouse chimeras derived from either genetically engineered ES cells or from diploid embryonic cells. In these cases there is complete segregation of the descendants of the ES cells. The tetraploid cells do not contribute at all to the formation of the embryo proper, but instead create the primitive endoderm derivatives and the trophectoderm (Naumann 2008). This method has also been successfully used for analyzing genes known to be heterozygous embryonic lethal as well as to rescue embryonic lethality caused by an additional maternally inherited X chromosome in the mouse (Carmeliet et al. 1996; Goto and Takagi 1998).

3.5.2 The Aneuploid Heteroploidies

Aneuploid heteroploidies are very common in the human species, where they represent up to 50 % of miscarriages. These heteroploidies are of two kinds, depending on whether the aberration results from the loss or from the gain of one (or more)

chromosome(s). *Monosomies* and *nullisomies*[5] characterize respectively the situation where a single chromosome or the two chromosomes of a given pair is (are) missing. *Trisomies* or *tetrasomies* represent the opposite situation, where the karyotype displays one or two extra copies of the same chromosome pair. All these numerical abnormalities result from errors occurring either during gametogenesis or at fertilization, and this explains why a karyotype with more than four copies of the same chromosome is virtually never observed in practice.

Aneuploid heteroploidies have all been modeled and studied systematically in the mouse because, as we will explain later they can be produced experimentally almost at will, with high frequency (see section referring to Robertsonian translocation later in this chapter). A number of conclusions have been drawn from these experiments and observations that we summarize here, but it must be kept in mind that the phenotype of the aneuploid heteroploidies, the trisomies in particular, depends upon the chromosome involved in the primary defect.

3.5.2.1 Nullisomies and Monosomies

As a rule, autosomal nullisomies and monosomies for any of the mouse autosomes are lethal in utero at an early stage. Nullisomics ($2n - 2$) are so severely affected that the condition is incompatible with egg segmentation: the conceptuses degenerate shortly after fertilization and are resorbed. Monosomics ($2n - 1$; symbol *Ms*) for an autosome can develop for a few hours, but most die prior to or during the implantation period and only rare survivors can be detected 6 days after fertilization. This early lethality probably indicates that, for many loci scattered over the autosomes, a 50 % reduction in gene expression is insufficient to assure normal embryonic development (Magnuson et al. 1985; Beechey and Searle 1988). Genomic imprinting is also likely involved.

3.5.2.2 Trisomies, Tetrasomies, Double Trisomies etc

Trisomies ($2n + 1$) result from chromosomal non-disjunction during gametogenesis or during the early stages of development. When the two chromosomes of a given pair, instead of migrating to the opposite poles during meiotic anaphase I or anaphase II, migrate to the same pole and go into the same daughter cell, this is called *non-disjunction*. This accident, which is relatively rare, results in a gamete being *disomic* while the complementary gamete is *nullisomic*. When such a disomic gamete fuses with a normal gamete (n chromosomes), this generates a trisomic embryo ($2n + 1$), while the nullisomic gamete when it fuses with a normal gamete results in a monosomic embryo ($2n - 1$). When chromosomal non-disjunction occurs during

[5] Sometimes called *nullosomies*.

embryonic development (i.e., in the somatic cells—during mitosis), the result is similar: a euploid mother cell, with 2n chromosomes, produces two daughter cells: one with a 2n + 1 complement and the other with a 2n − 1 complement. In this case, the embryo is a *mosaic*[6] of euploid and aneuploid cells. The aneuploid cells are sometimes counter-selected compared to the normal euploid cells, especially if they do not divide exactly at the same pace.

Trisomic embryos (symbol *Ts*) are often affected by severe specific defects. In the mouse, all individual trisomies, including those of the X chromosome, have been observed and studied in detail and, unlike in the case of the monosomic embryos, the morphology of affected trisomics is highly variable (Gropp et al. 1975). At one extreme are trisomics for chromosome 19 (symbol *Ts19*), the shortest autosome, which exhibit an almost normal morphogenesis up to 10 days in utero and then appear slightly delayed until birth. Some *Ts19* trisomic mice survive for a few days after birth, but many have a cleft palate and die. Mice trisomic for chromosome 12 (*Ts12*) also survive for quite a long time in utero, but all die at birth because they suffer from exencephaly. Mice trisomic for chromosome 14 and 16 also die at birth with rather characteristic pathological features. At the other extreme, trisomics for chromosome 2, 7, 8, and 15 have extremely severe phenotypes, with growth retardation and death occurring by the time of implantation or shortly after (Beechey and Searle 1988) (Fig. 3.9).

The viability of the trisomic conceptuses is not correlated with the size of the chromosome but probably with the density of genes, and this seems quite logical. In humans, for example, trisomy 21 is the only viable trisomy, probably because chromosome 21 is a small chromosome with only ~270 genes. Correlations between the origin of the extra chromosome (paternal or maternal) and the severity of the phenotype have not been clearly documented in the mouse but probably exist if one takes into account the phenomenon of parental imprinting (developed in Chap. 6).

Mouse primary trisomies, those involving a complete intact chromosome, are, unfortunately, not good models for studying human trisomies, for two reasons. First, the syntenic assortments of mouse and human genes on the different chromosomes are so different that no human chromosome has its faithful, complete, orthologous replica in the mouse species, and vice versa. Second, even if the mouse genes were distributed along the mouse chromosome exactly as they are in human, the phenotypes resulting from differences in gene dosage (3/2) in one species may be expressed differently in another. In other words, and for many reasons in addition to this one, mice are definitely not humans in reduction. However, analysis of mouse trisomies, in combination with human studies, sometimes provides a powerful system for understanding aneuploidy in both

[6] *Mosaics* refers to organisms whose cells have a different genetic makeup, although they are all derived from the same egg. Mice composed of both XO and XX cells, because one X was lost during development, for example, are mosaics. *Chimeras* are organisms whose cells do not have the same genetic makeup because they are derived from different embryonic cells. Mosaicism is natural, while chimerism is, in most instances, the result of experimental manipulation.

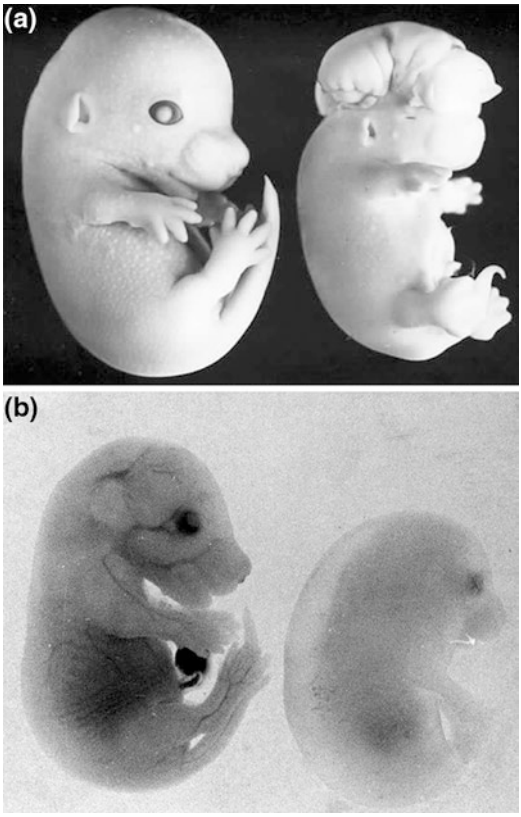


Fig. 3.9 *Trisomies*. **a** A mouse trisomic for chromosome 12 (day 18 p.c.) and its age-matched control. Note the exencephaly that is characteristic of this trisomy. (Courtesy of Dr. Heinz Winking, Medizinische Hochschule, Lübeck, Germany). **b** A mouse trisomic for chromosome 16 (day 15 p.c.) and its age-matched control. (Courtesy of Dr. Muriel Davisson, The Jackson Laboratory, Bar Harbor, Maine, USA). *Ts16* embryos are slightly retarded and edematous

species (Hernandez and Fisher 1999). In addition, mouse trisomies are excellent tools for studying the effect of variations in gene copy numbers.

Chimeric mice resulting from the association of trisomic cells with normal euploid cells (*Ts* ↔ 2n) have been produced experimentally and have revealed some interesting aspects of early tissue differentiation. It was repeatedly observed, for example, that in *Ts12* ↔ 2n chimeras, cells trisomic for chromosome 12 were able to participate in the formation of most tissues, including the ovary, but were never found in lymphocyte populations, presumably as a consequence of early negative selection in this particular cell lineage (Fundele et al. 1985). Other chimeric mice with a trisomic partner (Chr 16 for example), have been produced and have also been found to be fully viable, indicating that trisomic cells (at least some of them) can be successfully integrated in a developing chimeric embryo

and, accordingly, that they are not cell-lethal. This sort of experiment might still be used in the context of parental imprinting to analyze the consequences of gene dosage effects in some chromosomal regions (see Chap. 6).

Tetrasomies (2n + 2) and double trisomies (2n + 1 + 1) are extremely rare anomalies even when produced experimentally, and have not been studied in detail.

3.5.2.3 Aneuploidies of the Sex Chromosomes

Aneuploid heteroploidies concerning either the X or the Y chromosome have been frequently reported in laboratory mice for two reasons. First, these heteroploidies, unlike autosomal heteroploidies, are viable, even if the affected mice are often sterile. Second, mouse geneticists have excellent X-linked markers that allow the detection of sex chromosome numerical anomalies at a glance, simply by observing that the sex of the mouse does not match with the presumptive genotype for the sex chromosome (or *gonosomes*). The X-linked marker Tabby *(Ta)*, for example, has been extensively used in this context and has allowed the detection of XXY or XO individuals because of an unexpected striped or non-striped coat color.[7]

The viability of mice with extra X or Y chromosomes is no surprise if we consider that all X chromosomes but one in a karyotype are functionally inactivated (Chap. 6), and that the Y chromosome is a relatively gene-poor element.

In the mouse, monosomy of the X chromosome (39,X0) is compatible with normal survival and behavior. Female mice, unlike women who are affected by the homologous syndrome, Turner syndrome (45,X0), can breed but they produce many less X0 offspring than theoretically expected (i.e., 1/3). To explain this shortage, it has been suggested that, during gametogenesis, X0 females preferentially segregate the set of chromosomes lacking the X into the polar body (Morris 1968).[8] In contrast with X0 mice, women affected by Turner syndrome exhibit some phenotypic differences with normal women (short stature, sterility, etc.). Geneticists think that these differences are attributable to the functional haploidy of some X-linked genes that are not normally inactivated in human females. For example, the *SHOX* gene, which maps to the human pseudo-autosomal region (PAR), might be responsible for the short stature characteristic of Turner syndrome (X0) in human females, while its mouse ortholog is not X-linked and accordingly is not affected by the monosomy.

---

[7] Tabby (*Ta*) is an X-linked coat color and fur marker. $X^{Ta}X^{+}$ females are striped; $X^{Ta}Y$ males have a typical coat color with bare patches behind the ear, greasy fur, and a "sticky" tail. A *Ta*-striped male is then unexpected unless it is XXY. A female with a Tabby [*Ta*] phenotype is expected to be X0.

[8] The theoretically expected 25 % of mice with a 39,0Y karyotype die at a very early stage of development because of the X nullisomy. X0 females seldom produce more than 10 % X0 offspring and have a reduced stock of oocytes, resulting in a much shorter breeding period than normal XX females.

Mice with a XXY constitution (41,XXY) have been found but their frequency is very low (approximately 0.04 % among laboratory males and 0.08 % among wild-caught males in some populations). These males, equivalent to human Klinefelter syndrome, have a normal body mass and appearance, but significantly smaller testes than normal, and no visible germ cells (Cattanach 1961; Hauffe et al. 2010).

The XYY sex-chromosome constitution, which is relatively common in human, has also been described in the mouse (Cattanach and Pollard 1969). These males are sterile probably because of the combined deleterious effects of two Y chromosomes acting prior to meiosis, and pairing abnormalities leading to meiotic breakdown (Hunt and Eicher 1991).

3.6 Variations in Chromosome Structure

A great variety of structural variations has been described concerning the mouse karyotype. Some of these variations are relatively minor and are characterized, for example, by moderate uncoiling of chromatin in the peri-centromeric regions, imparting a more or less elongated appearance to one or a few chromosome pairs (Forejt 1973; Vig et al. 1994). In other instances, scientists observed that in a group of wild mice from a certain geographic area, a specific band is slightly enlarged, indicating local chromatin amplification (often described as *homogeneous staining regions*—HSR). These morphological variations of the karyotype have proved interesting for their cladistic value but they have in general little or no influence on the survival and behavior of the affected mice.

In contrast, some other structural changes resulting from chromosome breakages have more or less severe consequences, either on the survival of the embryo or on the fertility of the affected animals. We will review the most important of these structural abnormalities.

To understand the nature of these structural alterations, let us imagine that we are sitting in front of a panel on which the individual mouse chromosomes are displayed, assorted in pairs as they appear in the ideogram. Let us now imagine that we are handed a pair of scissors and requested to make cuts (one, two, three…) randomly in the chromosome arms, then to pick up the fragments and glue them back, also randomly, but always associating a telomeric fragment with a centromeric fragment, regardless of whether this was regenerating the original picture or whether this created a "recombinant" association. In fact, what we were demonstrating with the scissors analogy is exactly what happens in reality, either naturally or after exposition of the post-meiotic germ cells to X- or γ-rays or after injection of a chemical mutagen a few weeks before mating.

To complete this rather simple scenario we must make some additional comments:

- Because they are sticky, the ends of the broken chromosomes have a spontaneous tendency to reunite with other damaged ends (Schulz-Schaeffer 1980).

- Structural rearrangements resulting from a single break have severe consequences if the break is not repaired, because they split the chromosomes in two pieces: one that remains attached to the centromere and the other that segregates randomly during the mitotic or meiotic process. Such structural rearrangements are almost always lethal for the cell because, by definition, the genetic material is no longer distributed evenly after cell division. This explains why structural rearrangements resulting from a single break, paradoxically, are rare.
- Breaks can occur at any time, in the adult or in the embryo, in somatic or germ cells. If the alteration occurs in the germ line, it may interfere with the meiotic process, resulting in reduced fertility or even complete sterility of the affected individuals. Structural reshuffling occurring in the germ line may also lead to the production of gametes with an abnormal set of chromosomes (unbalanced gametes).

3.6.1 The Structural Rearrangements Resulting from a Single Break

These structural rearrangements are called *deletions*, or more precisely *terminal deletions* to distinguish them from the *interstitial deletions* that require two breaks (see below). These deletions are uncommon because they result in partial (or tertiary) monosomies for the telomeric fragment and accordingly have highly deleterious consequences. As mentioned above, a chromosome that has lost its telomere is unstable and is accordingly strongly counter-selected. Conversely, a fragment with no centromere (*acentric*) is rapidly lost when the cell divides.

3.6.2 The Structural Rearrangements Resulting from Two Breaks

These two breaks can occur on the same chromosome but, in most cases, they involve two different chromosomes.

3.6.2.1 The Structural Rearrangements Resulting from Two Breaks in Two Different Chromosomes

Reciprocal Translocations

Reciprocal or balanced translocations, as the name indicates, result from a reciprocal exchange between the telomeric ends of two non-homologous chromosomes. They

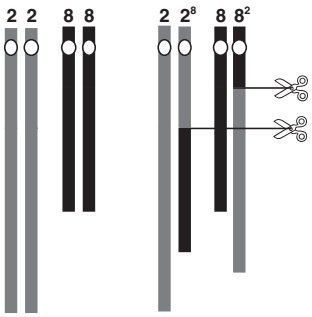


Fig. 3.10 *Reciprocal translocations*. Reciprocal translocations are the most common kind of chromosomal rearrangement (one break in two different chromosomes with reciprocal fusion). When heterozygous they cause reduced fertility (semi-sterility) or sometimes complete sterility, in one sex or the other. Some reciprocal translocations are viable and fertile when homozygous

are, by far, the most common form of structural rearrangement of the mouse karyotype and around 150 such translocations are listed in the Mouse Genome Database.[9]

The standard symbol used to define these reciprocal translocations is *T*. When the chromosomes involved in the translocation are identified, the symbol contains this information: *T(2;8)26H,* for example, is the 26th reciprocal translocation recorded at Harwell; it involves chromosomes 2 and 8. When the positions of the breakpoints relative to the G-banded karyotype are known, this can also be indicated by adding the band numbers after the corresponding chromosome numbers: the same *T26H* would then be designated *T(2H1;8A4)26H,* since the breakpoints are respectively in band H1 of Chr 2 and band A4 of Chr 8 (Beechey and Evans 1996) (Fig. 3.10).

Mice heterozygous for reciprocal translocations, in most cases, have no visible external phenotype,[10] which is in keeping with the fact that these structural rearrangements do not quantitatively alter the genetic makeup of the affected animals. However, some heterozygous mice are sterile in one sex or the other, sometimes in both.

Gametogenesis in mice heterozygous for a reciprocal translocation is always strongly perturbed, leading to the production of a high percentage of abnormal gametes. To explain this, we will consider the meiosis of a heterozygous mouse *T(2;8)26H/+* (de Boer and de Maar 1976) and as a simplification we will consider exclusively the chromosomes involved in the structural rearrangement.

[9] Chromosome 2 appears to be more frequently involved in reciprocal translocations than expected based on its size (27 occurrences). The reason for this bias is unknown.

[10] The reciprocal translocation *T26H* is one of the very few exceptions because it is associated with a coat-color change visible only in homozygous (*T26H*/*T26H*) animals. This change is probably a consequence of an alteration at the *Agouti* locus (Chr 2) generated by the structural rearrangement.

In these conditions, the genotype of the *T26H/+* mice could be symbolized as $2 + 8 + 2^8 + 8^2$, where 2 and 8 are the normal chromosomes, and 2^8 and 8^2 the recombinant (or translocated) chromosomes, the superscript being a reference to the origin of the telomeric segments. A *T26H/+* heterozygous mouse produces different kinds of gametes depending on the segregation of the chromosomes during meiosis. The first class is represented by the gametes with the intact chromosomes 2 and 8 ($2 + 8$). Another class corresponds to the translocated chromosomes 2^8 and 8^2. In these two chromosomes there is a redistribution of the genetic material, but there is no loss or gain and for this reason these gametes are called *balanced*. When a gamete carrying the two intact chromosomes ($2 + 8$) fuses with another normal gamete ($2 + 8$), this restores a fully normal mouse karyotype and the translocation is lost. When a gamete carrying ($2^8 + 8^2$) fuses with a normal gamete ($2 + 8$), this generates a mouse heterozygous for the translocation, like the heterozygous parent of the mating.

In addition to the gametes mentioned above, ($2 + 8$) and ($2^8 + 8^2$), mice heterozygous for the translocation *T(2;8)26H/+* also produce ($2^8 + 8$), ($2 + 8^2$), ($2 + 2^8$), and ($8^2 + 8$) gametes. All these gametes are unbalanced and when they fuse with a normal gamete, the resulting embryos are all unviable because some chromosomal segments are duplicated, while others are missing. For example, a conceptus with a karyotype ($2^8 + 8 + 2 + 8$) is, at the same time, monosomic for a (telomeric) piece of chromosome 2 and trisomic for a telomeric (distal) piece of chromosome 8. Geneticists say that conceptuses with such an unbalanced karyotype are *tertiary monosomic* for the telomeric segment of Chr 2 and *tertiary trisomic* for the telomeric portion of Chr 8. Embryos with this type of karyotype die in utero at an early stage, probably because of the tertiary monosomy. This wastage of conceptuses explains why the progenies of mice heterozygous for a reciprocal translocation (*T/+*) are always reduced in number. Mouse geneticists designate this reduced fertility by the term *semi-sterility*. Semi-sterility is a phenotype common to all reciprocal translocations that can be objectivized by looking at the uterine content of pregnant mice at day 13/16 of pregnancy, and observing that about 50 % of the implants are in the process of resorption (Fig. 3.11).

So far we have considered only the cases of crosses between mice heterozygous for a reciprocal translocation and normal mice (for example, *T26H/+* x *+/+*). However, intercrosses between mice heterozygous for the same translocation (for example, *T26H/+* x *T26H/+*) are also possible. In most instances, the offspring of such crosses are abnormal, with an unbalanced karyotype, and the living progenies are extremely reduced. However, two cases are noteworthy:

- The first case is when, for example, a ($2^8 + 8^2$) balanced gamete from one partner merges with another similar gamete ($2^8 + 8^2$) of the other partner. In this case, an embryo homozygous for the translocation ($2^8 + 2^8 + 8^2 + 8^2$) results. In the specific case of *T26(2;8)H*, the mouse is viable and fully fertile, but this is not a rule, and mice homozygous for some other reciprocal translocations are unviable or, less frequently, sterile. Carefully selected mice of this kind, with a

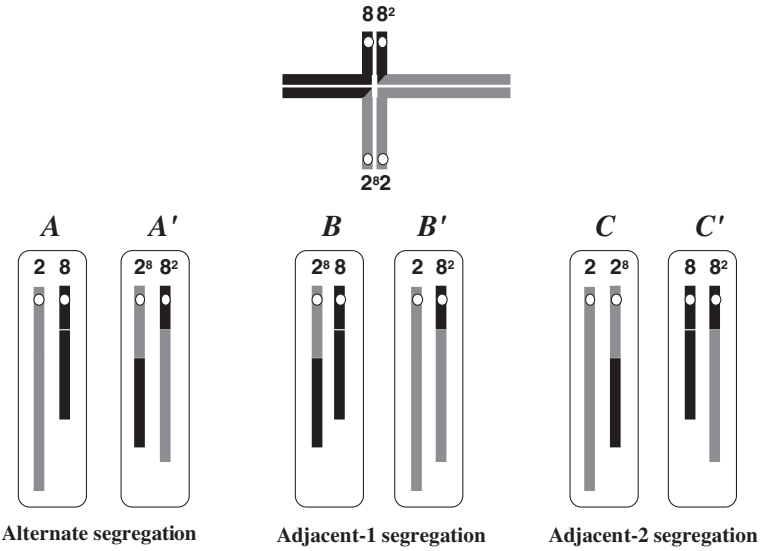


Fig. 3.11 *Meiosis in a mouse heterozygous for the reciprocal translocation T(2;8)26H*. Mice heterozygous for a reciprocal translocation (*T*/+), when fertile, have to form unusual meiotic structures in order for all chromosomal segments to pair with their homologous regions. Disjunction of the centromeres can be either alternate (*A* and *A′*), adjacent 1 (*B* and *B′*), or adjacent 2 (*C* and *C′*). Adjacent 2 is rare and is sometimes referred to as non-disjunction. Gametes resulting from alternate segregation are either normal (*A*) or balanced (*A′*), with one (and only one) copy of each gene. When such gametes merge with normal gametes of the opposite sex, this generates normal embryos, with either a normal karyotype or a karyotype with the reciprocal translocation. Gametes resulting from adjacent-1 segregation (*B* and *B′*) or adjacent-2 segregation (*C* and *C′*) are unbalanced and have either 0, 1, or 2 copies of each gene, depending on the chromosomal segment. When these gametes merge with a normal gamete, aneuploid embryos result. Reciprocal translocations have been very helpful for the establishment of genetic maps as well as for the analysis of the epigenetic mechanisms at work in parental imprinting

karyotype different from that of normal laboratory mice, have been used as a source of cells (for example, T or B lymphocytes) for performing transplantations or grafts because the different morphology of the chromosomes allows for tracking of the transplanted cells in the chimeric organism.

- Another interesting situation is when, unbalanced gametes with a complementary karyotype fuse together. To explain the situation, let us consider the case of another reciprocal translocation: *T(2;11)30H*. As in the case described above, mice heterozygous for this reciprocal translocation produce six kinds of gametes with the chromosomal constitution: $(2 + 11)$; $(2^{11} + 11^2)$; $(2 + 11^2)$; $(2^{11} + 11)$; $(11 + 11^2)$ and $(2^{11} + 2)$. However, when a non-balanced gamete with the constitution $(2^{11} + 2)$ fuses with the complementary gamete $(11 + 11^2)$ contributed by the sexual partner, this generates a euploid embryo $(2 + 11 + 2^{11} + 11^2)$, which is heterozygous for the translocation and has a balanced karyotype. However, in this embryo, the centromeric part of chromosome 2 comes from

the same parent, while the other parent contributes the centromeric segments of chromosome 11. This situation is also known in the human species and is designated as double non-disjunction or uniparental disomy (UpD).

Experiments focusing on the developmental potentialities of mouse embryos resulting from such double non-disjunctions have been achieved by scientists at the Harwell MRC Research Centre using a variety of reciprocal translocations and a variety of genetic markers, allowing the unambiguous identification of the parental origin of the chromosomal segments. The conclusion of these experiments is that, unexpectedly, euploid embryos resulting from complementary UpD are not always viable. Sometimes they are viable when the UpD is of maternal origin, but lethal when it is of paternal origin or vice versa, depending on the chromosomes involved. In some instances, the embryos are viable but smaller sized (or larger sized) than their littermates, depending on the crosses. This clearly indicates that the genetic contribution of one parent is not equivalent to the contribution of the other parent. We will come back extensively to this point in Chap. 6, which is devoted to epigenetics and parental imprinting.

All these peculiarities of reciprocal translocation have been extremely useful at several crucial steps in the development of mouse genetics. Because they disrupt the linkage relationships between the genes on the same chromosome and simultaneously create new linkage groups by associating genes that were originally non-linked, they were extensively used from the late 1970s to the early 1980s to assign each and every linkage group to a particular chromosome and to determine the position of the centromere for the linkage groups (Searle et al. 1971). The idea behind this strategy is that reciprocal translocations have, at the same time, a phenotype that one can observe with a microscope (i.e., a reshuffled karyotype) associated with semi-sterility and, when crossed, they exhibit new linkage relationships between their genes while the original ones are disrupted (see Chap. 4).[11]

Reciprocal translocations have also provided essential tools for the localization of genes associated with a variety of human cancers and hereditary diseases.

Another interesting point when crossing translocation carriers (*T/+*) is that, among the aneusomic embryos that are produced, some are, by accident, tertiary trisomics, i.e., trisomics for a small piece of chromosome or even for a complete "recombinant" autosome. The reciprocal translocation *T(14;15)6Ca*, for example, is characterized by a very unequal reciprocal exchange with a relatively long chromosome 14^{15} (actually longer than Chr 1) and a very small chromosome 15^{14} (shorter than Chr 19). When these mice are intercrossed they occasionally produce aneuploid conceptuses with an extra chromosome 15^{14}. These mice are viable, they are tertiary trisomics for a small piece of mouse Chr 15 (the centromeric end) and a small piece of Chr 14 (the telomeric end), and they have been used in very clever experiments to map the position of the centromeres in Chr 14 and 15 and to

[11] Experiments involving crosses between translocation carriers (*T/+*) are difficult to achieve because semi-sterility dramatically reduces progeny sizes. In addition, and as commented, some reciprocal translocation carriers are infertile, impeding many experiments.

clarify the cytological identification of linkage group III (Eicher and Green 1972). A procedure for genetic mapping, making use of the reciprocal translocations *T(X;7)1Ct* and *T(7;19)145H*, called the duplication-deficiency method, has also been reported (Eicher and Washburn 1978). Finally, and as we will explain further in this chapter, *Ts(17^{16})65Dn* tertiary trisomic mice have been used to model human trisomy 21, or Down syndrome.

Very unequal reciprocal translocations producing a long chromosome and a complementary very short chromosome are not common, but some have been described and are known as *tandems*. In some cases the very small chromosome is lost during cell division with no consequences, since, as we already mentioned, it consists mostly of heterochromatin. The consequence of this type of translocation is an irreversible reduction in the number of chromosome arms and centromeres. Such a tandem has been reported as a derivative of the reciprocal translocation *T(7;15)33Ad*, with breakpoints in bands 7A1 and 15F3. Outcrossing the original semi-sterile *T(7;15)* mice generated monosomic mice for the short marker 7^{15}. By intercrossing these mice, viable nullisomic progeny for chromosome 7^{15} were obtained that could be intercrossed to produce a breeding stock with 38 chromosomes (Schriever-Schwemmer and Adler 1993).

Robertsonian Translocations

Robertsonian translocations, named after the American biologist W. Robertson, who described them first, are the last case of structural rearrangement resulting from two breaks on different chromosomes. In fact, as for the tandem described above, they are a special kind of reciprocal translocation resulting in the fusion of two independent acrocentric chromosomes into a single unit that looks like a metacentric chromosome. For this reason, they are also designated *centric fusions* or *whole-arm translocations.*

The mechanism of formation of these reciprocal translocations is not completely elucidated and may not be the same in all cases. However, based on observations made after specific staining, cytogeneticists believe that most fusions, as for the tandems presented above, result from reciprocal translocation, with the chromosomal breakpoints being very close to the centromeres of two different acrocentric chromosomes (Fig. 3.12). Translocated chromosomes may then have either one or possibly two centromeres. If they have only one centromere, the structural rearrangement is irreversible whereas, if they have two, it is theoretically reversible. The complementary short chromosome, which in most cases contains nonessential genes and possibly no centromere, is usually lost after a few cell divisions, as in the case of tandems. This loss of centromeres is clearly an evolutionary mechanism leading to the reduction in chromosome number. However, if the karyotype is reduced (or virtually reduced) by one centromere, the fundamental number (the number of chromosome arms) remains the same and the global genetic information remains unaltered although it is distributed differently.

The common symbol for Robertsonian translocation is *Rb* and, when the arms (i.e., the acrocentric chromosomes) are identified, this is integrated into the

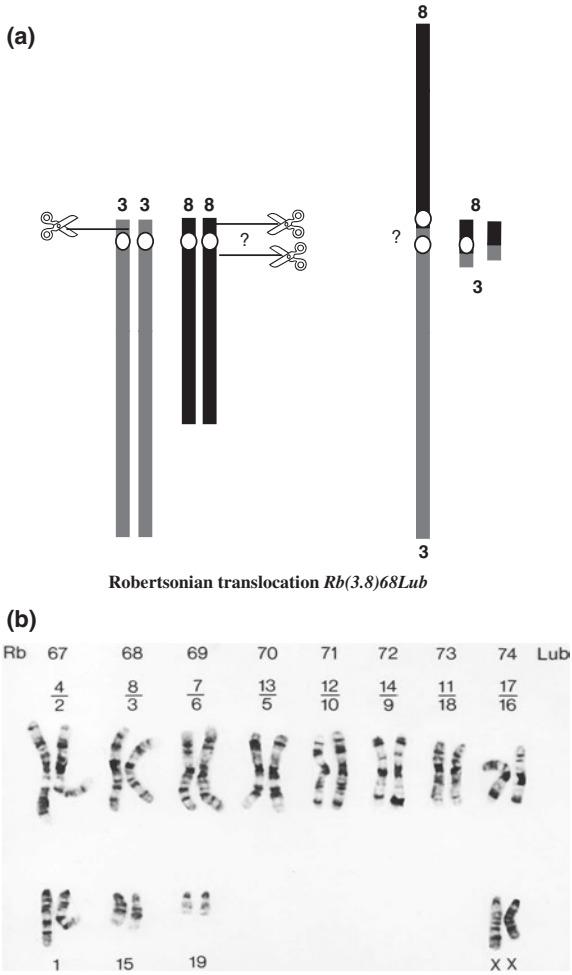


Fig. 3.12 *Robertsonian translocations*. Robertsonian translocations are an extreme case of reciprocal translocation. **a** In some cases (probably rare), one of the rearranged chromosomes has 2 centromeres and the other 0. In every case, when one of the two reciprocally rearranged chromosomes is very small, it is often lost. Such translocations (or centric fusions) are observed in many mammalian species and in particular in wild mice of the Alpine valleys. **b** The figure represents the G-banded karyoptype of a wild mouse trapped in Italy. Most chromosome pairs are involved in a Robertsonian translocation. Mice with such a karyotype are fully fertile, but their F1 with a fully acrocentric laboratory strain are nearly sterile due to a high frequency of meiotic non-disjunction. (Figure courtesy of Dr. Heinz Winking, Insitut für Biologie, Medizinische Universität zu Lübeck, Germany.)

symbol. *Rb(16.17)7Bnr*, for example, is a Robertsonian translocation resulting from the fusion of the acrocentric mouse chromosomes 16 and 17; this was the seventh Robertsonian translocation (*Rb7*) discovered by Alfred Gropp in a wild population of mice of the Poschiavo valley in Switzerland, and was bred in his laboratory in Bonn am Rhein (Bnr).

Robertsonian translocations are common polymorphisms in nature, especially in humans and in mice of the species *Mus m. domesticus*. In humans, only the acrocentric chromosomes (group D) are involved in these structural rearrangements; the three most important translocations (about one in a thousand newborns) being between chromosomes 13q and 21q, 14q and 21q, and 15q and 21q. *t(14q21q)* is frequently involved in Down syndrome and *t(13q21q)* in Patau syndrome.

In the mouse, Robertsonian translocations have been observed in laboratory populations (*Rb(9.19)163H* and *Rb(6.15)1Ald,* for example), and there are many isolates of wild animals in the Alpine valleys in Switzerland and Italy in which almost all chromosomes are metacentric. Isolates of this type are numerous, and many different associations of acrocentric chromosomes in centric fusions have been observed. It is even likely that the complete inventory for this kind of rearrangement has not yet been achieved in wild mice. Based on the observations made, and unlike what is observed in humans, there seems to be no restrictions on these acrocentric associations, even for the X chromosome. In contrast, the Y chromosome has never been found associated with any autosome in the form of a Robertsonian translocation.

Mice heterozygous for a Robertsonian translocation, unlike mice heterozygous for reciprocal translocations, are relatively fertile but exhibit a high percentage of chromosomal non-disjunctions during meiosis. This peculiarity has been exploited for the experimental production of trisomic and monosomic mice; we will briefly describe the experimental protocol.

When crossing a mouse homozygous for a Robertsonian translocation involving chromosome 17 (for example, *Rb(16.17)7Bnr*) with a mouse homozygous for another Robertsonian translocation involving the same chromosome 17 (for example, *Rb(8.17)Rma*), the F1 are heterozygous for both *Rb*s. If we ignore the chromosomes that are structurally identical in both partners of the cross, the genotype of these F1 can be symbolized as *Rb7Bnr+/+Rb8Rma* (i.e., 8 + 16 ↔ 17 + 8 ↔ 17 + 16). Mice with this type of karyotype (double heterozygous for *Rb* with monobrachial homology-17) are normal and fertile, but they generate a high percentage of unbalanced gametes due to non-disjunction at anaphase I. The gametes of these mice are either normal; for example, (8 ↔ 17 + 16) or (8 + 16 ↔ 17) or unbalanced; for example, (8 ↔ 17 + 16 ↔ 17) or (8 + 16). When these F1 mice are crossed with a normal mouse (8 + 8 + 16 + 16 + 17 + 17), up to 15 % of the embryos are trisomic for Chr 17 (*Ts17*) with a constitution (8 + 16 + 17 + 8 ↔ 17 + 16 ↔ 17). Similarly, the same percentage of monosomic embryos (*Ms17*) is also produced with a karyotype (8 + 16 + 17 + 8 + 16). Both these aneuploid embryos are easy to recognize based on the examination of their karyotypes (2 or 0 metacentric chromosomes). They both die at or shortly after implantation.

This situation is similar to the situation we reported for the reciprocal translocations, although, here, the resulting aneuploid offspring are *primary trisomics* or *monosomics* and not tertiary. Since there is a very large number of theoretically possible combinations for the production of double heterozygotes with monobrachial homology, all trisomic (or monosomic) animal models have been created and studied in the laboratory, for all chromosome pairs except X and Y (Gropp et al. 1975).

3.6.2.2 The Structural Rearrangements Resulting from Two Breaks in the Same Chromosome

Interstitial Deletions

As we mentioned earlier, when chromatids (or chromosomes) are broken, the cellular repair mechanisms are immediately activated and, depending on the breaks, the event may or may not result in a loss of genetic material. When there is a loss of genetic material, the structural alteration is called a *deletion* or *interstitial deletion* with the symbol *Del.*[12] When the deletion is cytologically visible in the karyotype, its designation takes this into account. *Del(5B1),* for example, designates a *deletion* of the band B1 of chromosome 5. Depending on their size, these deletions generally behave like dominant mutations with pleiotropic effects, less frequently as recessive ones. They are often lethal when homozygous.

Hundreds of deletions of this type were produced in Harwell (UK) and Oak Ridge and Argonne National Laboratory (USA) during the 20 years following World War II, while health physicists were studying the effects of X-rays, γ-rays, and neutrons on the genetic material of mammals. Many of these mutations have contributed to the development of the mouse linkage map, although some of them, because they are in fact small-sized chromosomal rearrangements rather than true point mutations, have been difficult to use, due to their suppressing effect on recombination leading to confounding results (compressions in the genetic maps) (Fig. 3.13).

The interest of deletions in genetics is well illustrated in the case of the analysis of the albino region (Chr 7-around the *Tyr* locus) by geneticists at the Albert Einstein College of Medicine (Gluecksohn-Waelsch 1979) and at Oak Ridge (Klebig et al. 1992). Deletions have been and still are of importance for mouse geneticists because, when they are numerous and overlapping, they allow the study, in great detail, of some regions of the genome and possibly the identification of new alleles after mutagenesis (see Chap. 7 regarding mutations and mutagenesis). Carefully selected deletions also allow the study of some regions of the mouse genome in the haploid state (see Chap. 6, devoted to the analysis of parental genomic imprinting). If all deletions that have been described in the mouse could be gathered into a single animal, they would make up about 1/4 of the total genome in the haploid state.

Finally, it is important to keep in mind that deletions frequently occur in vitro, in cell cultures, with (apparently) little or no consequences on cell growth and proliferation. For this reason, it is necessary to carefully and regularly check the karyotypes of embryonic cell lines (ES cell lines), making sure that they are always able to participate in the formation of viable chimeras with germinal transmission and to differentiate into all types of tissues. The presence of even small deletions

[12] A deletion (*Del*) is different from a *deficiency* (symbol *Df*) by its origin. Deficiency for a chromosome segment is generally associated with a *duplication* (*Dp*) of the same segment and results from the abnormal (unbalanced) segregation of a structurally rearranged chromosome.

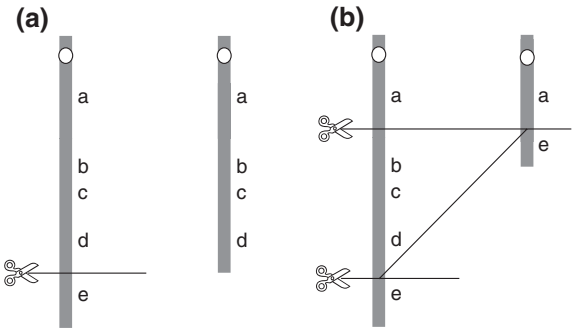


Fig. 3.13 *Schematic representation of chromosomal deletions*. **a** Terminal deletion (common in vitro and rare in vivo). **b** Interstitial deletion

in the genome of such ES cells may insidiously prevent their use for the production of genetically modified mice by homologous recombination in vitro.

Inversions

When a segment of a chromosome is isolated by two breaks, flipped over by 180°, and reintegrated into the same chromosome, the rearrangement is called an *inversion*. Rearrangements of this type, like reciprocal translocations, are relatively common events in all diploid species and participate in evolution at the chromosome level. They have an impact on the karyotype, because they generally alter the banding pattern characteristic of the affected chromosome.

When the chromosome segment generated by the two breaks includes the centromere, the inversion is said to be *pericentric*. When the centromere is outside the inverted segment, the inversion is called *paracentric*. In the normal laboratory mouse, pericentric inversions are virtually nonexistent, since the chromosomes are all acrocentrics and the centromeres are sub-terminal. The two types of inversions share the same symbol *In*, and when the affected chromosome is identified, here again the designation of the inversion takes this into account. *In(2)5Rk*, for example, designates an inversion of chromosome 2, which is the fifth chromosome anomaly of this kind collected by T.H. Roderick and colleagues.

Paracentric inversions have been induced experimentally in the mouse, with relatively high efficiency, by submitting male mice to either X- or γ-ray irradiation or by treating with alkylating mutagens, and by analyzing the offspring conceived in the following four weeks (Roderick and Hawes 1974).[13] Confirmation of the induction of an inversion in the offspring was based on the observation of so-called *anaphase bridges* in biopsies or sections of the seminiferous epithelium of the presumptive carriers (Fig. 3.14).

Paracentric inversions interfere with the normal meiotic process, because homologous chromatids cannot pair and come into close contact, as they generally

[13] In this case, the targeted cells were the late spermatids or spermatozoa.

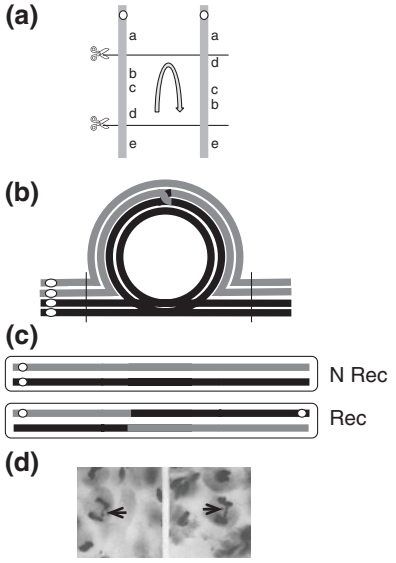


Fig. 3.14 *Paracentric inversions*. **a** Schematic representation of a paracentric inversion. **b** In mice heterozygous for a paracentric inversion, the homologous inverted chromatids form a loop that allows correct pairing. **c** When a crossing over occurs within the inverted segment (the loop), this generates acentric fragments (with no centromere) and the reciprocal dicentric fragments. **d** Since the two ends of a same dicentric chromatid are pulled apart, at the opposite poles of the cell, this results in an anaphase bridge that can be observed in a histological section of the testis (Figure courtesy of Dr. T. Roderick, The Jackson Laboratory, Bar Harbor, Maine, USA)

do during synapsis. To get around this problem, the inverted chromatids form a loop that allows the correct orientation for pairing, but when a recombination event (a crossing over) occurs in heterozygous mice within the inverted segment, this generates an acentric fragment (with no centromere) and a reciprocal dicentric fragment, with a centromere at both ends of the same chromatid. When these centromeres are pulled apart to the opposite poles during anaphase of the dividing cell, this causes an anaphase bridge that is often visible under the microscope (Torgasheva and Borodin 2001). In his initial description of the protocol for inducing inversions in the mouse, Roderick reported that several inversions induced by the alkylating agent tri-ethylene melamine (TEM) could yield up to 70 % of the cells exhibiting anaphase bridges (Roderick and Hawes 1974). In theory, the longer the inverted segment, the higher the observed frequency of anaphase bridges. In practice, however, this prediction is not confirmed, especially with the longer inversions, yielding a lower than expected percentage of anaphase bridges.

Since the chromosomes affected by paracentric inversions cannot easily pair with their normal homologous chromosomes during the pachytene stage of meiosis and, taking into account the fact that all crossing overs occurring within the inverted segments lead to a defective gamete, one can consider that

paracentric inversions act as virtual "crossing over suppressors" along the length of the inverted segment, and probably a little beyond the borders. For this reason, inversions are quite useful genetic tools for the recovery and maintenance of mutations in model organisms; in fact, they recreate a situation in the mouse analogous to the famous C*l*B (or Muller 5) condition designed by H. Muller for the induction and collection of X-linked mutations in the X chromosome of *Drosophila* (Muller et al. 1954).

In his experiments, Roderick noted that some hybrid males between the subspecies *Mus m. molossinus* and *Mus m. musculus* displayed a high frequency of first meiotic anaphase bridges, and sometimes double bridges, suggesting that the chromosomes of these subspecies may differ from those of the normal laboratory mouse by at least two paracentric inversions. This observation must be kept in mind because, if they indeed exist, such inversions may generate some difficulties in the analysis of mapping data when the subspecies *Mus m. molossinus* is a partner of the cross (as is often the case).

In contrast with deletions, inversions generally do not change the overall amount of the genetic material, and for this reason most of them are viable when homozygous. In some cases, one of the chromosome breaks is inside or in the close vicinity of a gene of essential function, and this sometimes generates a mutation with a visible phenotype. For example, the mouse mutation hairy ears (*Eh*-Chr 15), identified after neutron irradiation of post-meiotic germ cells due to the presence of a tuft of hair on the outside of the ear in heterozygotes, was later found to be at the breakpoint of an inversion spanning ~30 cM at the distal end of Chr 15. *Eh* is lethal at an early stage when homozygous (Davisson et al. 1990a). The case of *Eh* is uncommon and very few inversions have been found associated with a phenotype. In most instances, the inversions collected in specially designed experiments have been bred to homozygosity, indicating that the breakpoints are not frequently in essential regions (Katayama et al. 2009).

Inversions have some influence on gametogenesis of *In/+* carriers, since many of the gametes recombinant within the inverted segment are wasted. However, in most instances, this does not produce more than a slight reduction in fertility.

3.6.3 Complex Structural Rearrangements

The same game we began to play when making cuts in the karyotype and then re-associating the fragments in all possible positions could be pursued to the three-cut step, yielding increasingly complicated situations. In practice, very few structural rearrangements resulting from three chromosome breaks exist in the repositories, but at least two are worthy of comment. The first is the famous "Cattanach transposition" discovered in Harwell. The Cattanach transposition corresponds to the insertion (symbol *Is*) or transposition (symbol *Tp*) of a fragment of chromosome 7 in the middle of the mouse X chromosome (two breaks

in chromosome 7 and one in the middle part of the X chromosome). The full symbol of this rearrangement is (*Is(In7;X)1Ct* or *XCt*). Because the transposed/inserted segment contains the wild-type gene encoding tyrosinase (*Tyr*), the Cattanach transposition has been a useful tool for the study of X-chromosome inactivation (see Chap. 6). Albino mice heterozygous for *Is(In7;X)1Ct* appear "patchy," having a coat with pigmented patches on an otherwise albino background, depending on the active X-chromosome in the melanocytes. The other insertion is *Is(7;1)40H*, which corresponds to the insertion of part of Chr 7 into Chr 1. This insertion is male-sterile and has been used for the purpose of gene assignment.

3.6.4 Structural Rearrangements Created in Vitro

The great majority of the chromosomal rearrangements discussed in this chapter were found by chance, either among the offspring of mice submitted to mutagenic treatment resulting in chromosome breakages (X-rays, γ-rays, sometimes neutrons) or in wild mice. As we said, these chromosomal rearrangements have been very helpful, for example, for the assignment of mouse genes to specific chromosomes or for the orientation of linkage groups with respect to the centromeres (see Chap. 4). More recently, they have also been extremely useful for the discovery and genetic analysis of imprinted regions in the mouse (Chap. 6) and the discovery of homologies in the human species. Unfortunately, the enormous collection of chromosomal rearrangements has not been as useful as geneticists would have wished for the genetic analysis of trisomies for the simple reason that the mouse genes, even if they have high homologies (orthologies) with human genes at the molecular level, are nevertheless distributed very differently in the genome of each species, making it difficult to faithfully model a human trisomy. We have already noted that no mouse trisomy is viable for more than a few hours ab utero, while humans affected by Down syndrome can live for decades. Confronted with this intrinsic and unavoidable difficulty, geneticists have tried to develop better and more refined models by transposing the refined techniques they use for making alterations in the genome of ES cells to the field of cytogenetics. The cuts we made virtually in the former paragraphs, with a pair of scissors cutting the mouse chromosomes, can now be made extremely precisely (in fact, to the base pair) in the mouse genome by the molecular techniques of homologous recombination. This means that any kind of reciprocal translocation or inversion can now be "tailor-made". Similarly, extra pieces of mouse (or human) chromosomes can also be added to the mouse karyotype by trangenesis, simulating tertiary trisomies. We will summarize all these possibilities in the next section by presenting, as an example, the progress made in modeling Down syndrome.

3.7 Modeling Human Down Syndrome

Among the trisomies that affect the human species,[14] Down syndrome (DS—a trisomy of HSA21—or 47,XY + 21), is by far the most important for two reasons: (i) because of its relatively high frequency (approximately one newborn in 750 is affected) and (ii) because the syndrome is complex with highly variable and often severe pathologies including mental retardation, congenital heart defects, dysmorphic features, early-onset Alzheimer disease, increased risk of specific leukaemias, immunological deficiencies, and some other health problems.[15]

3.7.1 Mouse Trisomy 16: A Model of Down Syndrome

When human geneticists realized that Down syndrome (DS) was the consequence of an imbalance in gene dosage for some of the ~268 genes linked to human chromosome 21, and considering that a great number of the human genes on HSA21 have an orthologous copy on mouse chromosome 16 (MMU16), they had the logical idea to model DS by producing mice trisomic for this autosome (41,XY + 16) or (41,XX + 16).[16] Such mice can be easily produced, for example, by crossing mice double heterozygous for the Robertsonian translocations *Rb(16.17)7Bnr* and *Rb(6.16)9Rma* with normal laboratory mice (40,XX or 40,XY). Among the offspring of such crosses, most mice inherit only one of the two metacentric chromosomes, either *Rb7Bnr* or *Rb9Rma*, plus a complementary set of acrocentric chromosomes (Cox et al. 1984). However, in some instances (up to 10 %), nondisjunctions occur, the two metacentric chromosomes stay together to form a disomic gamete, and a trisomic offspring results when the gamete in question merges with a normal one. As expected, such trisomic mice exhibit some features characteristic of human trisomy 21 (edema, cardiac anomalies, etc.) but, unfortunately, the model had serious drawbacks (Epstein et al. 1985; Epstein 1990) (see Fig. 3.9b). First, the mice did not survive *ab utero* but died at a late stage of pregnancy, blocking some experiments, in particular behavioral tests. Second, and most importantly, although these mice were trisomic for the segment harboring the mouse genes orthologous to the genes on HSA21, they were disomic (i.e., normal) for some other genes of the same HSA21 that have homology with a segment of MMU10 or MMU17. Reciprocally, because MMU16 has syntenies with regions of

[14] Several autosomal primary trisomies have been described in the human species, but only three, trisomies for chromosome 13 (Patau syndrome), for chromosome 18 (Edwards syndrome), and for chromosome 21 (Down syndrome), affect live born children. Patau and Edwards syndromes are extremely severe. The relatively low gene density on chromosome 21 is consistent with the observation that trisomy 21 is one of the only viable human autosomal trisomies.

[15] HSA21 = abbreviation for *Homo sapiens* chromosome 21; MMU16 = abbreviation for *Mus musculus* Chr 16.

[16] This estimation of the number of genes on human chromosome 21 (HSA21) is from S. Scherer, *A short guide to the human genome*, Cold Spring Harbor Laboratory Press, 2008, p21.

HSA3, HSA8, and HSA16, many genes triplicated in Ts16 mice were not involved in the etiology of human DS. For these reasons, primary trisomics for MMU Chr 16 have been abandoned as models of DS.

3.7.2 *Ts(17^{16})65Dn: A Tertiary Trisomy Modeling Down Syndrome*

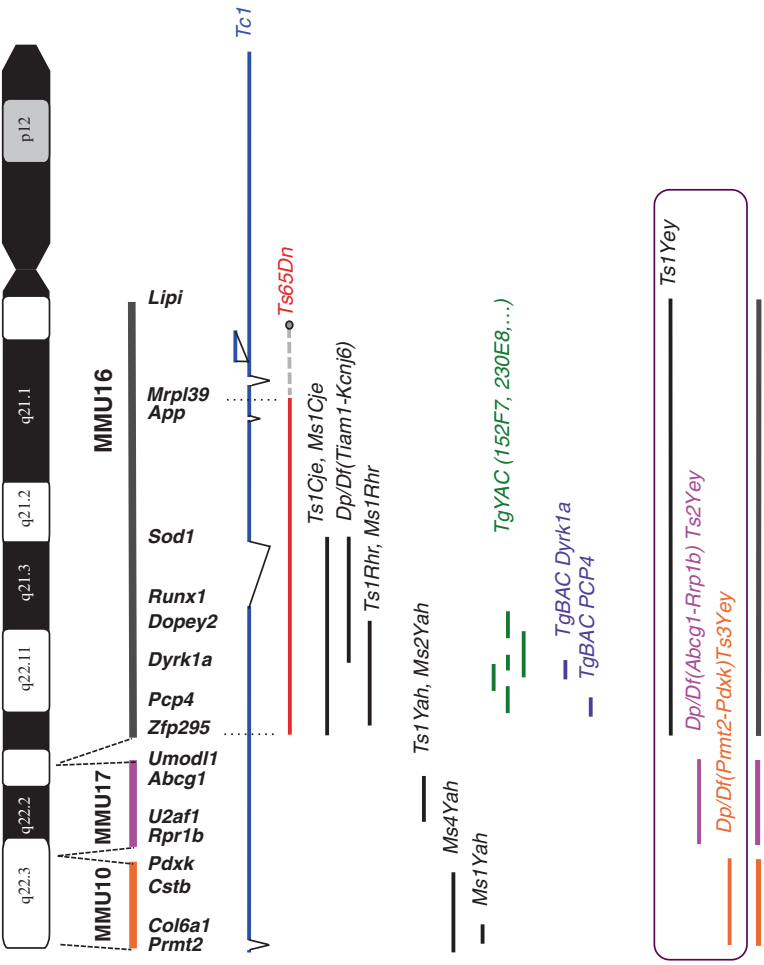
A more refined model of DS was developed by M.T. Davisson from The Jackson Laboratory, USA (Davisson et al. 1990a, 1993) and was extensively studied by R.H. Reeves and colleagues from Johns Hopkins University, Baltimore USA (Reeves et al. 1995). This model is commercially available under the name of *Ts(17^{16})65Dn*. These mice are tertiary trisomics, meaning that, in addition to the normal set of 40 chromosomes, they have in their karyotype an extra small chromosome resulting from a radiation-induced reciprocal translocation between Chr 16 and Chr 17 and comprising the centromere and proximal end of Chr 17 (~9.5 Mb) and the distal end of mouse Chr 16 (~34 Mb or ~100 genes, with an orthologous copy on HSA21). *Ts65Dn* mice survive to adulthood and exhibit many of the features of humans with Down syndrome. For example, they have spatial learning and memory defects and show some developmental delay. They also exhibit locomotor hyperactivity, lack of behavioral inhibition, and stereotypic behavior.

Ts65Dn mice are considered good models of DS, but they nonetheless also have some imperfections. The first and most important one is that only a segment of Chr 16 with orthology in the HSA21 segment 21q21-21q22.3 is triplicated in *T65Dn*. Another imperfection is that, here again, some genes that are triplicated in *Ts65Dn* mice are on Chr 17 and have no orthologous counterpart on HSA21. This variation in copy number probably interferes with the phenotype because some genes of MMU17, triplicated in the *Ts65Dn* mice, are known to play an important role in regulating neuronal functions.

In spite of these imperfections, *T65Dn* mice have the great advantage of exhibiting highly reproducible phenotypes, with clear similarities to DS, indicating that dosage imbalance for a gene or group of genes in the triplicated region definitely has a major contribution to this pathology. It is then likely that the corresponding dosage imbalance for the human orthologous copies of these genes also contribute to cognitive deficits in DS. These genes are part of the so-called *DS critical region*.

3.7.3 *Transgenic and Transchromosomic Models of Down Syndrome*

Many other models of DS have been developed over the past few years. Most of these models were created by pronuclear transgenesis (see Chap. 8) with cloned DNA (yeast artificial chromosomes or bacterial artificial chromosomes) of various



sizes from the relevant mouse chromosome, and assembled in the same individual by sexual reproduction. Depending on the genes in the transgenic segments, these models exhibited a variety of phenotypes more or less reminiscent of DS—and were considered "partial" models. Among these models, one results from the addition, in the same genome, after several rounds of crossing and selection, of three duplications (symbol *Dp*) of chromosomal regions of the mouse that are homologous to HSA21. These duplications are 2.3 Mb of MMU10 (*Dp(10)1Yev/+)* containing 41 genes, 1.7 Mb of MMU17 (*Dp(17)1Yev/+*) containing 19 genes, and 22.9 Mb of MMU16 (*Dp(16)1Yev/+*) containing 115 genes, all orthologous to HSA21 (Yu et al. 2010). The production of these models (around a dozen as of today) with copy number variation for regions homologous to HSA 21, has contributed to a better understanding of the individual influence of the different regions of human chromosome 21 on the brain alterations and a better definition of several DS critical regions.

◀ **Fig. 3.15** *Mouse models of Down syndrome*. This figure represents some of the different models of Down syndrome that have been described. On the *left-hand side* is a diagrammatic representation of human chromosome 21 (HSA21) with its three homologous regions in the mouse (MMU10, MMU17 and MMU16). *Tc1* is a diagrammatic representation of the transchromosomic mouse model indicating the segments of HSA21 that have been retained, lost, or duplicated. *Ts65Dn* is a tertiary trisomic mouse model with two segments of mouse chromosome, one of MMU16 (in *red*), containing a "Down syndrome critical region" of human chromosome segment 21q22 and another smaller one of MMU17 (*grey dotted line*). *Ts65Dn* mice are trisomic for ~13.4 Mb of the HSA21 syntenic region, containing approximately 99 orthologs of HSA21 genes, and are one of the most popular models of DS. Several other segmental trisomies for a shorter region of MMU16 (*T1Cje*, etc.) or transgenic strains for yeast artificial chromosomes or bacterial artificial chromosomes of MMU16 have also been published, but all these models exhibit a less severe phenotype than *T65Dn*. On the *right-hand side* is a model described by Yu et al. (2010) in which the regions of mouse chromosome 10, 16, and 17, syntenic to human chromosome 21 (*boxed*), have been duplicated in vitro, in ES cells, then assembled in the same genome by sexual reproduction. All the models represented here exhibit more or less faithfully some of the DS-related neurological defects and will certainly be very useful for understanding the cognitive disability associated with DS. Unfortunately, due to the complexity of the genetic interactions involved in DS cognitive phenotypes, it is likely that no mouse model will ever recapitulate the whole spectrum of intellectual disabilities observed in DS. (The *background* of this picture is courtesy of Dr. Yann Herault, Institut Clinique de la Souris, Strasbourg, France)

Finally, a truly original model has been created by genetic engineering in ES cells (see Chap. 8), leading to the production of a "transchromosomic" line that stably transmits a freely segregating, almost complete human chromosome 21 (HSA21) (O'Doherty et al. 2005). This model exhibits phenotypic alterations in behavior, synaptic plasticity, cerebellar neuronal number, heart development, and mandible size that relate to human DS.

Two comprehensive reviews of these models have been published explaining their peculiarities, advantages, and drawbacks (Herault et al. 2012; Rueda et al. 2012). None of these mouse models faithfully replicate human Down syndrome in all of its details, but many of them allow the identification of brain regions affected by homologous trisomies in the two species (Fig. 3.15).

3.8 Conclusions

The aim of this chapter was to provide an overview of the most important aspects of mouse cytogenetics, describing the most common chromosome aberrations and anomalies and their phenotypes or consequences on reproduction. We realize that this presentation is rather superficial and greatly simplified, and for this reason we provided references to the most important publications on the subject. The chromosome aberrations and anomalies we have listed here have proved to be invaluable tools for the establishment of the genetic map, for unraveling some aspects of genomic imprinting, and, more recently, for modeling Down syndrome. In the future, they may still prove useful in experimental contexts where gene dosage is an important issue.

Acknowledgments The authors are appreciative of the critical comments on the present chapter provided by Drs. Marie Geneviève Mattei (Hôpital de la Timone, Marseille) and Yann Herault (Institut Clinique de la Souris, Strasbourg).

References

Baron B, Métézeau P, Kiefer-Gachelin H, Goldberg ME (1990) Construction and characterization of a DNA library from mouse chromosomes 19 purified by flow cytometry. Biol Cell 69:1–8

Beechey CV, Evans EP (1996) Numerical variants and structural rearrangements. In: Lyon MF, Rastan S, Brown SDM (eds) Genetic variants and strains of the laboratory mouse, vol 2, 3rd edn. Oxford University Press, Oxford

Beechey CV, Searle AG (1988) Effects of zero to four copies of chromosome 15 on mouse embryonic development. Cytogenet Cell Genet 47:66–71

Blasco MA (2005) Mice with bad ends: mouse models for the study of telomeres and telomerase in cancer and aging. EMBO J 24:1095–1103

Carmeliet P, Ferreira V, Breier G, Pollefeyt S, Kieckens L, Gertsenstein M, Fahrig M, Vandenhoeck A, Harpal K, Eberhardt C, Declercq C, Pawling J, Moons L, Collen D, Risau W, Nagy A (1996) Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. Nature 380:435–439

Caspersson T, Lindsten J, Lomakka G, Moller A, Zech L (1972) The use of fluorescence techniques for the recognition of mammalian chromosomes and chromosome regions. Int Rev Exp Pathol 11:1–72

Caspersson T, Lindsten J, Lomakka G, Wallman H, Zech L (1970) Rapid identification of human chromosomes by TV-techniques. Exp Cell Res 63:477–479

Cattanach BM (1961) XXY mice. Genet Res 2:156–158

Cattanach BM, Pollard CE (1969) An XYY sex-chromosome constitution in the mouse. Cytogenetics 8:80–86

Cox DR, Smith SA, Epstein LB, Epstein CJ (1984) Mouse trisomy 16 as an animal model of human trisomy 21 (Down syndrome): production of viable trisomy 16 diploid mouse chimeras. Dev Biol 101:416–424

Cox EK (1926) The chromosomes of the house mouse. J Morphol Physiol 43:45–54

Davisson MT, Roderick TH, Akeson EC, Hawes NL, Sweet HO (1990a) The hairy ears (Eh) mutation is closely associated with a chromosomal rearrangement in mouse chromosome 15. Genet Res 56:167–178

Davisson MT, Schmidt C, Akeson EC (1990b) Segmental trisomy of murine chromosome 16: A new model system for studying Down syndrome. Prog Clin Biol Res 360:263–280

Davisson MT, Schmidt C, Reeves RH, Irving NG, Akeson EC, Harris BS, Bronson RT (1993) Segmental trisomy as a mouse model for Down syndrome. Prog Clin Biol Res 384:117–133

de Boer P, de Maar PHMD (1976) A histological study of embryonic death caused by heterozygosity for the T26H reciprocal mouse translocation. J Embryol Exp Morph 35:595–606

Eicher EM, Green MC (1972) The T6 translocation in the mouse: its use in trisomy mapping, centromere localization, and cytological identification of linkage group 3. Genetics 71:621–632

Eicher EM, Washburn LL (1978) Assignment of genes to regions of mouse chromosomes. Proc Natl Acad Sci U S A 75:946–950

Epstein CJ, Cox DR, Epstein LB (1985) Mouse trisomy 16: an animal model of human trisomy 21 (Down syndrome). Ann N Y Acad Sci 450:157–168

Epstein C (1990) Genetic control of survival of murine trisomy 16 fetuses. Teratology 42:571–580

Forejt J (1973) Centromeric heterochromatin polymorphism in the house mouse. Evidence from inbred strains and natural populations. Chromosoma 43:187–201

Fundele R, Jägerbauer EM, Kolbus U, Winking H, Gropp A (1985) Viability of trisomy 12 cells in mouse chimaeras. Wilhelm Roux's Arch Dev Biol 194:178–180
Gluecksohn-Waelsch S (1979) Genetic control of morphogenetic and biochemical differentiation: lethal albino deletions in the mouse. Cell 16:225–237
Goto Y, Takagi N (1998) Tetraploid embryos rescue embryonic lethality caused by an additional maternally inherited X chromosome in the mouse. Development 125:3353–3363
Green J, Ried T (eds) (2011) Genetically Engineered mice for cancer research: design, analysis, pathways, validation and pre-clinical testing. Springer
Gropp A, Kolbus U, Giers D (1975) Systematic approach to the study of trisomy in the mouse II. Cytogenet Cell Genet 14:42–62
Hauffe HC, Giménez MD, Garagna S, Searle JB (2010) First wild XXY house mice. Chromosome Res 18:599–604
Herault Y, Duchon A, Velot E, Maréchal D, Brault V (2012) The in vivo Down syndrome genomic library in mouse. Prog Brain Res 197:169–197
Hernandez D, Fisher EM (1999) Mouse autosomal trisomy: two's company, three's a crowd. Trends Genet 15:241–247
Hunt PA, Eicher EM (1991) Fertile male mice with three sex chromosomes: evidence that infertility in XYY male mice is an effect of two Y chromosomes. Chromosoma 100:293–299
Katayama K, Miyamoto S, Furuno A, Akiyama K, Takahashi S, Suzuki H, Tsuji T, Kunieda T (2009) Characterization of the chromosomal inversion associated with the *Koa* mutation in the mouse revealed the cause of skeletal abnormalities. BMC Genet 10:60. doi:10.1186/1471-2156-10-60
Kaufman MH, Lee KKH, Speirs S (1989) Influence of diandric and digynic triploid genotypes on early mouse embryogenesis. Development 105:137–145
Kim SH, Parrinello S, Kim J, Campisi J (2003) *Mus musculus* and *M. spretus* homologues of the human telomere-associated protein TIN2. Genomics 81:422–432
Klebig ML, Kwon BS, Rinchik EM (1992) Physical analysis of murine albino deletions that disrupt liver-specific gene regulation or mesoderm development. Mamm Genome 2:51–63
Kubiak JZ, Tarkowski AK (1985) Electrofusion of mouse blastomeres. Exp Cell Res 157:561–566
Leeb M, Wutz A (2011) Derivation of haploid embryonic stem cells from mouse embryos. Nature 479:131–134
Liyanage M, Coleman A, du Manoir S, Veldman T, McCormack S, Dickson RB, Barlow C, Wynshaw-Boris A, Janz S, Wienberg J, Ferguson-Smith MA, Schröck E, Ried T (1996) Multicolour spectral karyotyping of mouse chromosomes. Nat Genet 14:312–315
Magnuson T, Debrot S, Dimpfl J, Zweig A, Zamora T, Epstein CJ (1985) The early lethality of autosomal monosomy in the mouse. J Exp Zool 236:353–360
Miller OJ, Miller DA (1975) Cytogenetics of the mouse. Annu Rev Genet 9:285–303
Miller OJ, Miller DA, Kouri RE, Allderdice PW, Dev VG, Grewal MS, Hutton JJ (1971) Identification of the mouse karyotype by quinacrine fluorescence, and tentative assignment of seven linkage groups. Proc Natl Acad Sci U S A 68:1530–1533
Morris T (1968) The XO and OY chromosome constitutions in the mouse. Genet Res 12:125–137
Muller HJ, Herskowitz IH, Abrahamson S, Oster II (1954) A nonlinear relation between x-ray dose and recovered lethal mutations in drosophila. Genetics 39:741–749
Naumann R (2008) Production of tetraploid mouse embryos by electrofusion. Biocompare article, Monday, 04 Aug 2008
Nesbitt MN, Francke U (1973) A system of nomenclature for band patterns of mouse chromosomes. Chromosoma 41:145–158
Niemierko A (1981) Postimplantation development of CB-induced triploid mouse embryos. J Embryol Exp Morphol 66:81–89
O'Doherty A, Ruf S, Mulligan C, Hildreth V, Errington ML, Cooke S, Sesay A, Modino S, Vanes L, Hernandez D, Linehan JM, Sharpe PT, Brandner S, Bliss TV, Henderson DJ, Nizetic D, Tybulewicz VL, Fisher EM (2005) An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. Science 309:2033–2037

Reeves RH, Irving NG, Moran TH, Wohn A, Kitt C, Sisodia SS, Schmidt C, Bronson RT, Davisson MT (1995) A mouse model for Down syndrome exhibits learning and behaviour deficits. Nat Genet 11:177–184

Roderick TH, Hawes NL (1974) Nineteen paracentric chromosomal inversions in mice. Genetics 76:109–117

Rueda N, Flórez J, Martínez-Cué C (2012) Mouse models of Down syndrome as a tool to unravel the causes of mental disabilities. Neural Plast 2012:584071, Article ID 584071. doi:10.1155/2012/584071

Sahin E, De Pinho RA (2010) Linking functional decline of telomeres, mitochondria and stem cells during ageing. Nature 464:520–528

Sawyer JR, Moore MM, Hozier JC (1987) High resolution G-banded chromosomes of the mouse. Chromosoma 95:350–358

Schriever-Schwemmer G, Adler ID (1993) A mouse stock with 38 chromosomes derived from the reciprocal translocation T(7;15)33Ad. Cytogenet Cell Genet 64:122–127

Schulz-Schaeffer J (1980) Cytogenetics plants—animals—humans. Springer, New York

Searle AG, Ford CE, Beechey CV (1971) Meiotic disjunction in mouse translocations and the determination of centromere position. Genet Res 18:215–235

Sumner AT, Evans HJ, Buckland RA (1971) New technique for distinguishing between human chromosomes. Nat New Biol 232:31–32

Torgasheva AA, Borodin PM (2001) Synapsis and recombination in inversion heterozygotes. Biochem Soc Trans 38:1676–1680

Uhlmann F (2013) Open questions: chromosome condensation—why does a chromosome look like a chromosome? BMC Biol 11:9

Vig BK, Latour D, Frankovich J (1994) Dissociation of minor satellite from the centromere in mouse. J Cell Sci 107:3091–3095

Yu T, Li Z, Jia Z, Clapcote SJ, Liu C, Li S, Asrar S, Pao A, Chen R, Fan N, Carattini-Rivera S, Bechard AR, Spring S, Henkelman RM, Stoica G, Matsui S, Nowak NJ, Roder JC, Chen C, Bradley A, Yu YE (2010) A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. Hum Mol Genet 19:2780–2791

Zhang S, Teng Y (2013) Powering mammalian genetic screens with mouse haploid embryonic stem cells. Mutat Res 741–742:44–50

Zhu L, Hathcock KS, Hande P, Lansdorp PM, Seldin MF, Hodes RJ (1998) Telomere length regulation in mice is linked to a novel chromosome locus. Proc Natl Acad Sci U S A 95:8648–8653

Chapter 4
Gene Mapping

4.1 Introduction

Now that the sequence of the mouse genome is completely known, the position of any gene of the species can be accurately and rapidly established by searching the appropriate database. In this context, a chapter devoted to gene mapping and genetic maps might appear somewhat outdated, not to say useless. However, we thought that it might be interesting to reconsider this subject for at least three reasons. The first is that gene mapping has been a major component of the activities of mouse geneticists during most of the twentieth century; it is then interesting, if only from a historical point of view, to briefly describe the techniques and methods that have made the genetic map of the mouse the richest and most documented map of all mammals, including humans, for nearly 50 years. The second reason is more fundamental and refers to the many mutations that occur spontaneously in the breeding nuclei of inbred strains or those that are induced by mutagenic agents. All these mutations are initially characterized by an abnormal phenotype and some of them may appear of potential interest, for example, as models of human diseases. However, annotating and characterizing all these mutations requires that they be first carefully located on a chromosome and analyzed at the molecular level, when relevant. Finally, and as we will discuss in Chap. 10, understanding the determinism and mechanisms at work in the transmission and expression of quantitative traits requires that the genetic determinants of these traits be accurately identified, and this always begins with a mapping experiment.

4.1.1 The Discovery of Linkage Groups: A Historical Perspective

After the initial observations of Sutton, Boveri, and Morgan (reported in Chap. 3), it was recognized that the genes in the mouse nuclear genome were all physically

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_4

associated with one or other of the 19 autosomes or X–Y chromosomes.[1] Under these conditions, it was implicitly accepted that two (or more) genes, once on the same chromosome, would have a tendency to remain associated or "linked" together in the same parental association, generation after generation, while all other genes, those that were each on different chromosomes, would segregate independently. It was also known that this physical association between genes on the same chromosome was not permanent or absolute, since recombination (crossing-over or chiasmata) was regularly observed during meiotic prophase, when homologous chromatids paired. It was then logical to guess that the probability for such an event to occur between any two loci would depend upon the physical distance between the loci in question. To express it differently, two genes that are distant by only a few kilobases (kb) on the same chromosome would almost always segregate together because the probability for a crossing-over to occur and split the association is very low. In contrast, two other genes that are, for example, 50 Mb apart will have a much greater chance of being separated by a recombination event. As a consequence of this principle, when two genes are relatively close to each other on the same chromosome they no longer segregate independently and this, by definition, generates distortions from the expected Mendelian proportions. This fundamental concept of genetic linkage was introduced in 1906, by Bateson and Punnett, after a series of experiments on the inheritance of comb shape in chickens (Bateson and Punnett 1906).

This situation, which nowadays may appear obvious, was not so simple to unravel. In 1904, Darbishire, lecturer in Genetics at the University of Edinburgh, reported the results of crosses involving two recessive alleles, chinchilla (c^{ch}, now $Tyr^{c\text{-}ch}$) and pink-eyed dilution (*p*, symbolized *e* in the original publication and now $Oca2^{p}$) and concluded that, in this particular cross, Mendel's law did not apply (Darbishire 1904). This statement was in contradiction with Cuénot's observations, reported in Chap. 1, demonstrating that Mendel's law indeed applied to mammals (Cuénot 1902). A few years later, Haldane et al. (1915), re-examining Darbishire's observations, interpreted the latter as resulting from "reduplication" or *linkage*, as we would now say. This linkage between two loci was the first to be reported in any vertebrate species. Darbishire's results were replicated in many other laboratories, including in Haldane's, using the same chinchilla allele ($Tyr^{c\text{-}ch}$) or another recessive allele (extreme dilution $Tyr^{c\text{-}e}$) at the same albino locus (*C* or *Tyr*).[2] The discovery of this linkage indicated that the loci for *c* and *p* were presumably at some distance on the same chromosome, but the chromosome in question was not known and the two genes were simply considered as the first members of "linkage group I".

After this first observation, many other coat color and phenotypic markers were used in a variety of crosses, and more linkages were progressively discovered. In 1927, Gates (1927) could not obtain any mouse that was simultaneously dilute and

[1] The structure of the mitochondrial genome (or mtDNA) is discussed in Chap. 5; here we consider only the genes in the nuclear genome.

[2] In this type of cross, the classical recessive allele Tyr^{c} (albino) cannot be used because it has an epistatic interaction with pink-eyed dilution, affecting eye and coat color, which makes phenotyping difficult.

short-eared (i.e., *d/d*; *se/se*) among an enormous F2 progeny of 1,312 offspring. He got three categories of offspring: 321 dilutes (*d/d* and +/*se*?); 653 wild-type (+/*d*? and +/*se*?),[3] and 338 short-eared (*se/se* and +/*d*?), compatible with a 1:2:1 ratio, and accordingly he concluded in "*absolute linkage*".[4]

Two years later Lord and Gates (1929) reported that *shaker-1*, a mouse mutation with a head-shaking behavior (the old symbol was *sh-1* and is now *Myo7a*^{sh1}), was also linked to both the *c* and *p* loci. This was confirmed by Grüneberg (1935) after analysis of a large cross (1,144 mice). After these observations the notion of "linkage group" was clearly established, with linkage group (LG) I now consisting of three genes: *p*, *c*, and *sh-1*. Linkage group II had only two genes, dilute (*d*) and short-ears (*se*), but, progressively, other genes were also reported linked to these genes. At the end of World War II, 30 genes were found to be members of one or the other of the 10 identified LGs. They were 72 in 1955, with 15 LGs, 162 in 1965 (Eicher 1981) and 217 in 1971 with a set of 20 LGs.[5] Twenty, by the way, was exactly the theoretically expected maximum number of LGs, since there are 19 autosomes and one X and one Y chromosome in the mouse genome.[6, 7] At this point several questions could then be addressed: (i) what is the gene order in a given LG?; (ii) to which chromosome should a given LG be assigned?; and (iii) at what end of a LG is the centromere of the corresponding chromosome? The last two questions, although they have required a lot of work and a lot of skillfully designed crosses, were solved in less than 4 years (1971–1975). It took much longer to finally integrate the individual mapping data into a single consensus map. We will briefly review the different steps of the establishment of the mouse genetic (or linkage or meiotic) map.

4.2 From Linkage Groups to Genetic Maps

4.2.1 Detecting Linkage and Measuring the Distances Between Loci

For a gene to be mapped, a fundamental prerequisite is that at least two allelic forms of this gene be available and that these two allelic forms be involved in the same cross with other allelic forms at the neighboring loci on the same linkage

[3] +/*d*? indicates that the actual genotype of the mouse is not known. It may be either +/+ or +/*d*.

[4] Reading the original publications is sometimes difficult due to the use of a nomenclature system different from the one in use nowadays. The *dilute* locus, for example, was designated "density" with two alleles: *D* and *d*. Nowadays, the same gene is symbolized *Myo5a*^{d}.

[5] A review by Dr. Eva Eicher from The Jackson Laboratory is a rich source of information concerning the historical aspects of mouse gene mapping and the progressive development of the genetic map in this species (Eicher 1981).

[6] It is now established that there are very few genes on the Y chromosome.

[7] The presumption that the twenty LGs identified at that time were each located on different chromosomes was shown to be wrong. In fact, a few genes were still mis-assigned and one LG was not yet identified.

group (or chromosome). Nowadays, as we will discuss further, the situation is different and much simplified because many genes are characterized at the molecular (DNA) level and the genotype can be considered to be merged with the phenotype. Let us, however, stay for another few pages at the pre-molecular era to outline as simply as possible the basic principles for the detection of genetic linkage.

The notion of genetic linkage, as we said, means that the parental allelic associations have a tendency to remain unchanged in the successive progenies unless a recombination event occurs to split the association in question. To make this clear, let us imagine two genes at two different loci on the same chromosome: *A* and *B*. Alleles *A* and *B* are fully dominant over the recessive forms *a* and *b*, and we will assume that all four alleles are fully viable and fully penetrant. The male parent is homozygous for the dominant allele *A* and for the recessive allele *b* while the other parent, the female, is homozygous for the recessive allele *a*, and homozygous for the dominant allele *B*. As a convention, such genotypes will be denoted *Ab/Ab* for the male and *aB/aB* for the female.[8] The F1 offspring of this cross will all have the same genotype *Ab/aB* and, when intercrossed, these F1 will produce an F2 in which one expects to get a variety of genotypes. If these genes are distant although still on the same chromosome, non-parental (or recombinant) allelic associations will be common and we will observe four phenotypes: [AB], [Ab], [aB], [ab], with proportions close to the expected Mendelian proportions 9/16, 3/16, 3/16, 1/16 (Table 4.1a–c).

If the two loci are tightly linked (as in the case reported above for the dilute (*d*) and short ears (*se*) loci on LG II), then only three phenotypic classes will be observed: [AB], [Ab], and [aB], with the proportions 1/2, 1/4, and 1/4, while the phenotype resulting from two recombinant chromosomes [ab] will virtually never occur. Finally, Table 4.1c represents an intermediate situation in which one third (1/3) of the gametes are recombinant and the rest (2/3) non-recombinant. In this case we will still observe the expected four phenotypic classes, but the one resulting from the fusion of two recombinant chromosomes will be less frequent.

In the example we just described, we mated a male with the genotype *Ab/Ab* and a female with the genotype *aB/aB*, then we mated the F1 (*Ab/aB* × *Ab/aB*) to produce the F2 offspring. This sort of cross, an *intercross* or F2 cross, is common because in most instances the two recessive alleles, *a* and *b*, were discovered independently and in different populations or strains, and accordingly there is a very low probability of finding them associated on the same chromosome (*ab/ab*) just by chance. However, if such a genotype occurs, either spontaneously or among the offspring of a cross, then it would be possible to cross a mouse with the genotype *AB/AB* with mice with the genotype *ab/ab*. The F1 mice would then have the genetic constitution *AB/ab* and the F2 would be of the same kind as above, although the recombinant genotypes would be different and the phenotypic proportions would also be different in the case of linkage. Finally, if the mating can be set up between a male *AB/ab* and a female *ab/ab*, then the situation would be much more advantageous for the detection of linkage and somewhat simpler to

[8] If the alleles at the *A* and *B* locus were not linked or if the linkage was not known, the symbols for the genotypes would be: *A/A*; *b/b* for the male and *a/a*; *B/B* for the female.

Table 4.1 The upper part of this table (**a**, **b** and **c**) represents the expected proportions of the different phenotypes [AB], [Ab], [aB] or [ab], in the progeny of an intercross for two alleles *A-a* and *B-b* in repulsion (*Ab/aB*)

a		male gamete			
		Ab - 1/4	aB - 1/4	AB - 1/4	ab - 1/4
female gamete	Ab - 1/4	Ab/Ab [Ab] 1/16	aB/Ab [AB] 1/16	AB/Ab [AB] 1/16	ab/Ab [Ab] 1/16
	aB - 1/4	Ab/aB [AB] 1/16	aB/aB [aB] 1/16	AB/aB [AB] 1/16	ab/aB [aB] 1/16
	AB - 1/4	Ab/AB [AB] 1/16	aB/AB [AB] 1/16	AB/AB [AB] 1/16	ab/AB [AB] 1/16
	ab - 1/4	Ab/ab [Ab] 1/16	aB/ab [aB] 1/16	AB/ab [AB] 1/16	ab/ab [ab] 1/16
		[AB] 9/16	[Ab] 3/16	[aB] 3/16	[ab] 1/16

b		male gamete			
		Ab - 1/2	aB - 1/2	AB - 0	ab - 0
female gamete	Ab - 1/2	Ab/Ab [Ab] 1/4	aB/Ab [AB] 1/4	AB/Ab [AB] 0	ab/Ab [Ab] 0
	aB - 1/2	Ab/aB [AB] 1/4	aB/aB [aB] 1/4	AB/aB [AB] 0	ab/aB [aB] 0
	AB - 0	Ab/AB [AB] 0	aB/AB [AB] 0	AB/AB [AB] 0	ab/AB [AB] 0
	ab - 0	Ab/ab [Ab] 0	aB/ab [aB] 0	AB/ab [AB] 0	ab/ab [ab] 0
		[AB] 1/2	[Ab] 1/4	[aB] 1/4	[ab] 0

c		male gamete			
		Ab - 1/3	aB - 1/3	AB - 1/6	ab - 1/6
female gamete	Ab - 1/3	Ab/Ab [Ab] 1/9	aB/Ab [AB] 1/9	AB/Ab [AB] 1/18	ab/Ab [Ab] 1/18
	aB - 1/3	Ab/aB [AB] 1/9	aB/aB [aB] 1/9	AB/aB [AB] 1/18	ab/aB [aB] 1/18
	AB - 1/6	Ab/AB [AB] 1/18	aB/AB [AB] 1/18	AB/AB [AB] 1/36	ab/AB [AB] 1/36
	ab - 1/6	Ab/ab [Ab] 1/18	aB/ab [aB] 1/18	AB/ab [AB] 1/36	ab/ab [ab] 1/36
		[AB] 19/36	[Ab] 8/36	[aB] 8/36	[ab] 1/36

d		male gamete			
		AB - 1/4	*ab* - 1/4	*Ab* - 1/4	*aB* - 1/4
f g	*ab* - 1/1	*AB/ab* [AB] 1/4	*ab/ab* [ab] 1/4	*Ab/ab* [Ab] 1/4	*aB/ab* [aB] 1/4

e		male gamete			
		AB - 1/2	*ab* - 1/2	*Ab* = 0	*aB* - 0
f g	*ab* - 1/1	*AB/ab* [AB] 1/2	*ab/ab* [ab] 1/2	*Ab/ab* [Ab] 0	*aB/ab* [aB] 0

f		male gamete			
		AB - 1/3	*ab* - 1/3	*Ab* - 1/6	*aB* - 1/6
f g	*ab* - 1/1	*AB/ab* [AB] 1/3	*ab/ab* [ab] 1/3	*Ab/ab* [Ab] 1/6	*aB/ab* [aB] 1/6

a when the two loci are not or only very loosely linked, in this case 50 % of the gametes are recombinant; **b** when they are tightly linked, in this case there are virtually no recombinant gametes; and **c** when they are moderately linked, in this case one third of the gametes are recombinant and the other 2/3 are non-recombinant. The proportion of mice homozygous for the two recessive alleles (*ab/ab*) varies from 1/16 (absence of linkage) to 0 (absolute or complete linkage). The lower part of the table (**d**, **e** and **f**) specifies the different genotypes (and phenotypes) of the offspring of a testcross or backcross: *AB/ab* x *ab/ab*. The female partner, being homozygous for both the *a* and *b* alleles, produces only one type of gamete. This sort of cross allows one to assess easily if the loci *A* and *B* are linked or not. If they are linked (**e** and **f**), the proportions of the different genotypes are different from the Mendelian proportions and vary from 0 to a value statistically lower than 1/4 (or 25 %). Computing the recombination frequency allows one to measure the distance between loci *A* and *B*. In the cases where the recessive alleles *a* and *b* are not fully penetrant or if some genotypes are unviable, the phenotypic class may be under-represented and a correction must then be applied.

analyze, since the recombination events occurring in the heterozygous parent (the male in our case) would all be informative and easy to score just by looking at the phenotypes of the offspring. The expected theoretical proportions in this case would be: 1/4 [AB], 1/4 [Ab], 1/4 [aB], and 1/4 [ab] and any deviation from these proportions would be suggestive of linkage, the fraction of recombinant genotypes being [Ab] + [aB]/total number of offspring (Table 4.1d–f).

Crosses of this second type, with a male *AB/ab* and a female *ab/ab* or the reverse, are called *testcrosses* (because one can test for linkage directly by counting the different categories of phenotypes in the offspring population) or *backcrosses* because the cross involves the F1 and a partner whose genotype is like the one of the *ab/ab* parent.

The genetic constitution *AB/ab* is designated *double heterozygotes* in *coupling*, while the reciprocal genotypic constitution *Ab/aB* is called *double heterozygotes* in *repulsion.*[9]

Intercrosses or F2 have, at the same time, a drawback and an advantage. The drawback is because the detection of linkage requires the phenotyping of more mice than in a testcross to reach the same level of significance. In a testcross one has to check whether the four phenotypic classes observed in the progeny match with the theoretically expected proportions 1/4, 1/4, 1/4, and 1/4, while in an F2 one has to check whether the four phenotypic classes match the classical proportions 9/16, 3/16, 3/16, and 1/16. On the other hand, the intercrosses or F2 have the advantage that, by genotyping a single individual, in fact we analyze the results of two meioses—one in each parent—not just a single one. This advantage will become more obvious when discussing molecular or co-dominant markers.

Once a situation of linkage is established between any two loci, a second step must then be envisaged: assessing the strength of this linkage; in other words, estimating the distance between the two loci. To explain this point we will take another simple example: we will mate a male mouse heterozygous in coupling for two dominant and two recessive alleles at the *C* and *D* loci (genotype *CD/cd*) with many *cd/cd* female mice. This is a simple testcross—it will produce four classes of offspring: *CD/cd* [CD], *cd/cd* [cd], *Cd/cd* [Cd], and *cD/cd* [cD]. Mice with a [CD] or [cd] phenotype are those resulting from non-recombinant gametes produced by the male, while the other mice, those with a [Cd] or [cD] phenotype, result from recombinant gametes. If we breed 317 offspring and observe, for example, 38 mice with either a [Cd] or [cD] phenotype, the ratio of recombinant offspring would be 38/317 = 0.11987 (i.e., 11.98 %). This result is an estimation of the actual linkage between *C* and *D* and if we were to repeat the experiment many times, we would get different results fluctuating around the value mentioned above. It is also intuitive that if we had raised ten times more offspring (3,170 instead of only 317) we would have obtained a more reliable estimate of the actual recombination frequency between the *C* and *D* loci. This is basic statistics, and formulas are available to compute the most likely estimate of the recombination

[9] When one of the mutant alleles (*M*) is dominant over wild type (+): the phase is *+M/am* for coupling and *aM/+m* for repulsion. In other words, the dominant alleles are associated on the same chromosome when in coupling.

frequency based on the number of mice scored in the progeny. For example, in the case reported above the confidence interval (at the 5 % risk level) for the recombination frequency is given by the formula:

$$p = p_o \pm 1.96\sqrt{\frac{p_o q_o}{N}}$$

where p_o is the observed ratio of recombinant offspring (in our example it is $38/317 = 0.11987$), $q_o = (1 - p_o) = 0.88013$, and N is the total number of mice scored for the phenotype (317 in our case). This yields 0.11987 ± 0.0356 (11.9 ± 3.5 %).

With a number of 3,170 offspring (and, say, 387 recombinants), the recombination frequency would be 0.1220 ± 0.0113 (12.2 ± 1.1 %) at the same 95 % confidence level.

Many computer programs are available for analyzing these data and in most instances it is sufficient to define the sort of cross, to introduce the number of mice in the different phenotypic classes and the required degree of confidence for the test (in general 5 %, sometimes 1 %), and the result is computed almost instantly.[10]

For geneticists, a 1 % recombination frequency is equivalent to one unit of map distance and is expressed as one centiMorgan (cM), a unit coined in homage to T.H. Morgan.[11]

It is important to clearly distinguish between recombination fractions and genetic distances. Recombination fractions are not additive measures. In fact, if one considers three loci *A*, *B*, and *C*, with *B* located between *A* and *C*, the recombination fraction between A and C is often less than the mere sum of the distances between *A* and *B*, and between *B* and *C*. This is due to the fact that any individual recombinant between *A* and *B* and between *B* and *C* will be counted as non-recombinant between *A* and *C*. To overcome this non-additivity of recombination fractions, a number of geneticists, starting with J.B.S. Haldane, designed mapping functions that convert recombination fractions into genetic distances, which are additive. The various mapping functions make different assumptions on the independence of crossing-overs that occur close to one another. Effectively, when a crossing-over occurs at a given position, there is a highly reduced probability that another one will occur nearby. This phenomenon is called *interference*, and varies in strength and extent between species. It is quite strong in the mouse, so that double recombinants are rare and recombination fractions can be reliably added for distances lower than 25 cM.

[10] Many of these software programs are listed at: http://www.jurgott.org/linkage/ListSoftware.html. The most popular are MAPMAKER, MAPMANAGER and GENE LINK.

[11] When the computed genetic distances are short or very short (<3 cM), it is recommended to express them with the lower and upper limits of the exact 95 % confidence interval calculated from the binomial distribution, as they appear in Table D5 and D6 (pp. 303–304) of Silver's book *Mouse Genetics*: *concepts and application*, Oxford University, 1995. This textbook is freely available at the Mouse Genome Informatics website.

When the loci *C* and *D* are distant the recombination fraction reaches 50 %, meaning that every other gamete is recombinant, and the two loci in question segregate independently, just as if they were located on different chromosomes. This point is confirmed by experimental data.

Another important point to consider is related to the sex of the heterozygous progenitor. In the testcross reported above, the male was the heterozygous partner with the genetic constitution *CD/cd*, while females were all homozygous for the recessive allele at this locus (*cd/cd*), producing only one sort of gamete (*cd*). However, it is now well established that the genetic distances computed from male meiosis are generally not the same as those estimated from female meiosis (Petkov et al. 2007). If our cross had been set up with the heterogametic sex (the male) being *cd/cd*, and the homogametic sex (the female) *CD/cd*, the estimation of the genetic distance would certainly have been different. On average, the recombination rate is higher in the homogametic sex than in the heterogametic sex. In some chromosome regions, however, such as subtelomeric or imprinted regions, this ratio is inverted. This means that a computed genetic distance is no more than an estimation of the actual physical distance, but these distances become more and more accurate as data from independent crosses, involving the two sexes, accumulate.

The genetic distances computed in cM by definition have an equivalent in DNA units (i.e., in kb or Mb), even if we know that this equivalent is not uniform as a consequence of the variations we just mentioned and others that will be discussed later. This equivalence has been estimated in the mouse by several means (to be discussed elsewhere), and a rough estimate is that one cM of mouse genome equals ~1.70 Mb of genomic DNA on average.[12, 13] Once the distances between loci have been established, *linkage maps* can then be constructed, in which loci are positioned according to the distance between them. However, before we can draw a map involving several genes, we must order all these genes linearly on their respective LGs.

4.2.2 Ordering the Genes

As we said above, 20 linkage groups were characterized in the late 1960s, meaning that, within each group, any individual gene had been found linked to at least one other gene of the same group. However, no information was available concerning the order of all the genes in the same LG. If it was established, for example, that the three genes *F*, *G*, and *H* were all members of the same linkage group we could not decide about the order of the three loci: it could be *F*, *G*, *H* or *F*, *H*, *G* or *G*, *F*, *H*.

---

[12] The physical (or DNA) size of the genome is estimated to be 2.7 Gb, and the mouse genetic (meiotic) map is estimated to span ~1,600 cM.

[13] This equivalence between cM and kb/Mb applies to the mouse only. In the human species, 1 cM is equivalent to ~0.7–1 Mb of DNA.

Knowing the genetic distances between the genes taken by pairs sometimes gives an indication. If, for example, we find that genes *F*, *G*, and *H* have the respective distances *F*–8 cM–*G*; *G*–11 cM–*H;* and *F*–18 cM–*H*, it is likely that the *G* locus is between the two loci for *F* and *H*. But such an ideal situation is not common, and distances such as *F*–0.5 cM–*G*; *G*–16 cM–*H;* and *F*–15 cM–*H* do not allow the loci to be ordered. Taking into account the experimental errors that always occur, the order is ambiguous and may be *F–G–H* or *G–F–H*. In such a situation, the best procedure is to set up what geneticists call a *three-point backcross.*

Such a cross consists of mating, for example, a male heterozygous at the three loci *F*, *G*, and *H* with a group of females, all homozygous for the recessive alleles at the same loci: *F G H/f g h* × *f g h/f g h*, and then to carefully phenotype all the offspring at all three loci. Since we have no a priori knowledge of the actual gene order, we will tentatively choose the alphabetical order *F–G–H* and classify the eight phenotypic groups.

Let us then assume that the experimental results are as follows, for a total of 1,078 mice:

[FGH]	= 360	non-recombinant
[fgh]	= 372	non-recombinant
[Fgh]	= 66	recombinant between F and G loci
[fGH]	= 68	recombinant between F and G loci
[FGh]	= 13	recombinant between G and H loci
[fgH]	= 17	recombinant between G and H loci
[fGh]	= 93	recombinants between F and G and G and H (double recombinant)
[FgH]	= 89	recombinants between F and G and G and H (double recombinant)

In this case, we can immediately observe that, taken two by two, the different reciprocal classes of phenotypes (for example, [fGH] and [Fgh] or [FGh] and [fgH]) are of the same order of magnitude (66/68 for the former group, 93/89 for the latter).

We also note that the phenotypic classes [fGh] and [FgH], resulting from gametes with a (supposed) double crossing-over, are quite large, whereas we would have expected them to be less frequent (two recombination events). This simply means that we were wrong when guessing a priori the order to be the alphabetical order *F–G–H*. In fact, if we modify the order and put the *H* locus between the *F* and *G* loci (the order now being *F–H–G*), then the phenotypic classes [FhG] (13 mice) and [fHg] (17 mice) are both double recombinant and less numerous.

We now have coherent data and the right order, and can then undertake the computation of the distances between the loci by applying the methods described above for only two loci. We will find that there are $66 + 68 + 13 + 17 = 164$ mice out of a total of 1,078 whose genotype is recombinant between the *F* and *H* loci, yielding an estimated distance of 15.2 ± 2.1 cM (at the 5 % risk). For the other two loci we will find that $93 + 89 + 13 + 17 = 212$ mice are recombinant between the *H* and *G* loci on a total of 1,078, giving an estimated distance of 19.6 ± 2.4 cM (at the 5 % risk).

With these distances and gene order, we can now draw a map with the locus *F* being at one extremity, the locus *G* at the other extremity, and the locus *H* in between. At this point, however, we have no idea of the centromere position but we know that it must be before F or after G given that the mouse chromosomes are all acrocentric (the centromere is near one end) (Fig. 4.1a, b).

When displaying the results of the three-point testcross in the example described above, we noted that the phenotypic classes were coherent in number and concluded that the three recessive alleles *f*, *g*, and *h* were fully viable and fully expressed in the homozygotes. Indeed, if we compute the number of mice with either a [f..] or [.g.] or [..h] phenotype, the three classes will be close to 50 %. Unfortunately, this is not a common situation. Frequently, one phenotypic class of offspring exhibits a shortage because a recessive allele is less viable than its dominant counterpart or because some mice are misclassified as a consequence of a non-expressed or mildly expressed phenotype (lack of expressivity or penetrance). This is quite common, for example, with mutations affecting either the eye or the skeleton, and when this

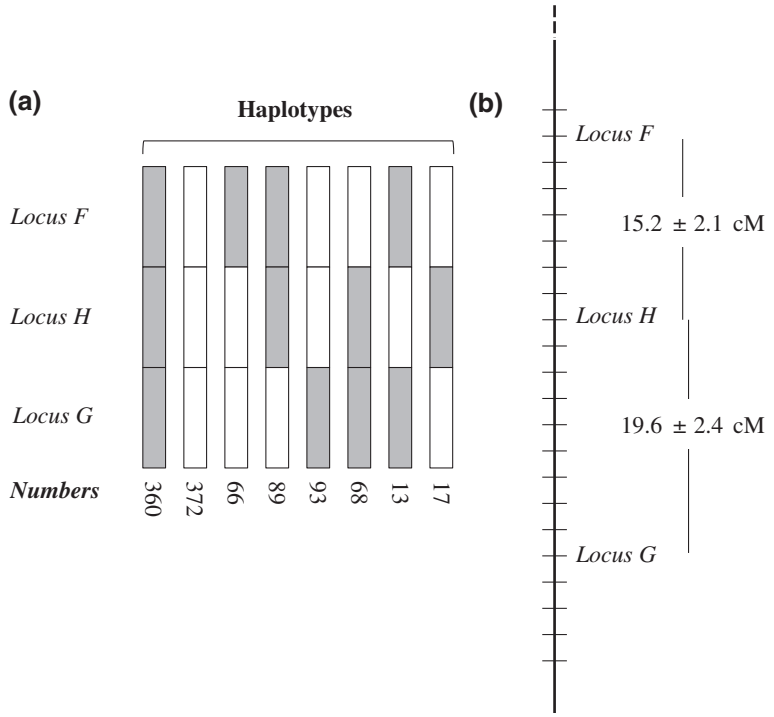


Fig. 4.1 *A three-point testcross*. A three-point backcross or testcross (*FHG/fhg* × *fhg/fhg*) allows one to order the genes linearly and estimate the genetic distances between the different loci. **a** The different haplotypes (*grey* heterozygous; *white* homozygous) for markers *F/f*, *H/h*, and *G/g*, respectively. **b** The genetic map drawn from the analysis of the genotyping data. The class of offspring whose phenotype is the least numerous corresponds to a double cross-over. Positioning the centromere requires other experiments

occurs the two classes of reciprocal recombinants are not equivalent. In this case, a correction must be made before computing the distances, and geneticists generally consider that the class with the largest number of recombinant offspring is probably a better estimate for computing the recombination frequency. Not taking this into account would result in an underestimation of the genetic distances.

Both interference and expressivity or penetrance are parameters that must be seriously taken into account in mapping experiments. Corrections for interference are extensively discussed in the book by Silver (1995) that is freely available at the Mouse Genome Informatics website. We strongly recommend this excellent book to readers with interest in issues related to genetic mapping.[14]

In the experiment reported above, we ordered the genes by setting up a three-point backcross. In fact, the strategy can equally apply to more genes, and setting up a four-point or five-point backcross is perfectly conceivable. Such crosses would be fairly difficult to prepare with classical phenotypic markers, but they are commonly performed with molecular markers, as we will explain later.[15]

Finally, one must remember that when linearly ordering the genes of one species it is always interesting to have a look at the situation in other species where the gene is known. As we will discuss in Chap. 5, homologies in the linear arrangement of genes are highly preserved across the different mammalian species, especially when they are short. In this case, geneticists speak of *homology of synteny* or *conservation of synteny.*

4.2.3 Establishing a Correspondence Between LGs and Chromosomes

Once the genes in a linkage group are linearly ordered, the next step consists of identifying the chromosome that encompasses the LG in question. Nowadays, the problem would be easily solved by selecting a molecular probe corresponding to one of the genes of the LG in question, labeling it with a fluorescent dye and performing fluorescence in situ hybridization (FISH) on a chromosome preparation as described in Chap. 3.[16] The localization could be confirmed by using another

[14] http://www.informatics.jax.org/silverbook/frames/frame9-1.shtml.

[15] Preparing a four-point backcross involving traditional or classical genetic markers, i.e., those that are scored by scrutinizing the mice one after the other, requires a lot of crosses because the markers in question are almost always in independent stocks or strains and first have to be gathered in the same stock by sexual reproduction.

[16] Most of the genes used for mapping in the past are now cloned and their DNA sequence is known. For LG I, for example, *Tyr* is the gene encoding tyrosinase and *Oca2* (oculocutaneous albinism II) is a gene encoding a transmembrane transporter essential for normal pigmentation. Both genes are involved in the production of melanin, and both are cloned and sequenced. The two genes can then be considered under two aspects: either as a mouse with a specific coat color, or as fluorescent dots on a mouse chromosome (see Chap. 3). In this particular case, it is chromosome 7 (bearing the whole of LG I).

cloned gene, belonging to the same LG as the previous one, and then labeling it with another fluorescent dye for another in situ hybridization. In this latter case, not only would the location of the LG be confirmed but, in favorable conditions, the position of the centromere would also be revealed (see Chap. 3, Fig. 3.6).

Unfortunately, FISH was not available when the LGs of the mouse were awaiting allocation to a specific chromosome, and in these conditions geneticists had to use other strategies. The basic principle of these strategies consisted of matching some morphological characteristics observed at the karyotype level [for example, the size of the chromosome, the specific G-banding pattern, the presence of a secondary constriction or of a negative-staining heteropycnotic (NHR) region, etc.] with the mapping data collected independently and concerning the reciprocal translocation breakpoints. The reciprocal translocations, because they result from two breaks on two different chromosomes with reciprocal exchange of the telomeric segment, split the LGs that are usually associated with the normal chromosomes and create two new ones. In these conditions, the chromosomal breakpoints themselves can be mapped like ordinary genetic markers since they have a phenotype (semi-sterility of the heterozygotes as a consequence of chromosome reshuffling) and a genotype (they induce differences in linkage). For example, from karyotypic observations performed at Harwell, the translocation *Rb163H* was demonstrated to involve a very short autosome, in fact the shortest of the karyotype (Chr 19), and a medium-sized acrocentric partner (Chr 9). From crosses involving phenotypic markers the same *Rb163H* fusion was found to involve both LG II and LG XII, meaning that one of the two LGs was on chromosome 19. The ambiguity was resolved after the observation that the reciprocal translocation *T145H* involved the same short autosome as *Rb163* (Chr 19), and the two linkage groups LG I and LG XII. The logical conclusion was that chromosome 19 encompassed LG XII (Lyon 1969; Eicher 1971).

Other similar experiments were conducted in different labs, involving different chromosomal rearrangements (mainly reciprocal translocations) and the same methodology.[17]

Other strategies have also been used for establishing a correspondence between LGs and chromosomes, for example by analysis of the expression profile of a specific gene in cells of teratocarcinomas from the LT/Sv strain, or in cells of trisomic embryos, or simply by looking for small morphological differences (Davisson et al. 1976; Eicher 1978; Eicher and Washburn 1978).

[17] The breakpoint of the reciprocal translocation *T(2;8)26H* on chromosome 2 is within the *Agouti* locus and inactivates this gene. Mice homozygous for the translocation, which are easy to identify by analysis of the karyotype, are also non-agouti (with a black coat color). This observation (and others) allowed it to be established that LG V (including the *Agouti* locus) was on Chr 2.

4.2.4 Positioning the Centromere

Once the different loci on a given linkage group were unambiguously ordered and assigned to a specific chromosome, the last ambiguity to be considered was the position of the centromere. This was relatively difficult to resolve if we recall that, in the mouse species, at least in the laboratory strains, all chromosomes are acrocentric with no genetic markers available on the short arm, at least at that time. In these conditions, positioning the centromere was impossible just by using classical mapping techniques. To achieve this, geneticists used a variety of different approaches that are discussed in detail in the book by Silver (1995). Among these approaches, the most popular and efficient consisted of using Robertsonian

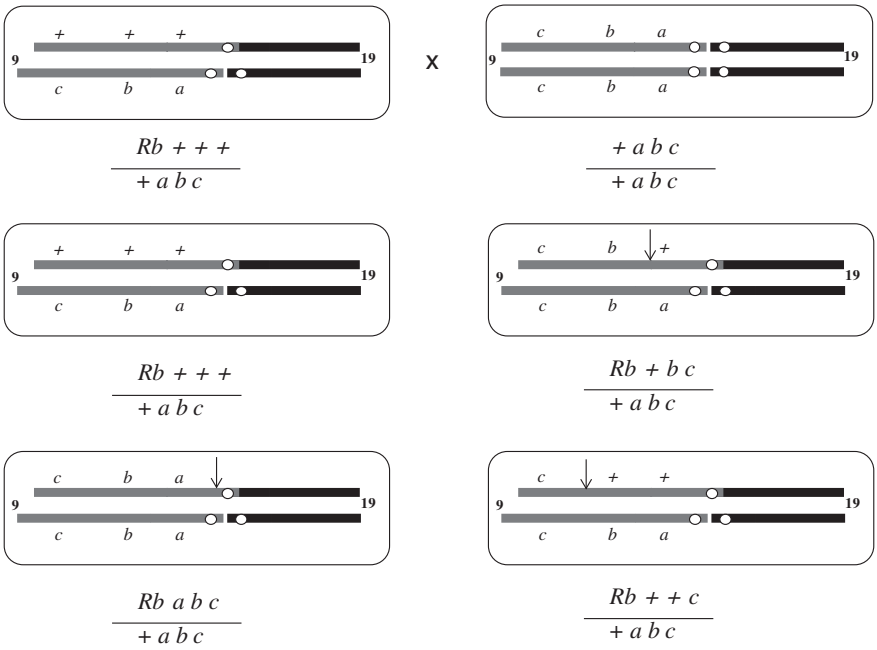


Fig. 4.2 *Positioning the centromere*. Robertsonian translocations have been useful tools for positioning the centromere once a linkage group is assigned to a specific chromosome. The figure represents a backcross between a mouse heterozygous for several recessive markers of chromosome 9 (*a*, *b* and *c*) and the Robertsonian translocation *Rb(9.19)163H* (*left*), and a mouse homozygous for the same three recessive markers and a normal karyotype (*right*). Analysis of the phenotypes of the offspring at the *a*, *b*, and *c* loci and for the presence/absence of the centric fusion *Rb163* in the karyotype indicates the relative position of the centromere. If mice homozygous for the *c* marker and heterozygous for the *Rb163* metacentric chromosome are more frequently observed than mice homozygous for the *a* marker and heterozygous for the same *Rb163* marker, this means that the genetic distance to the centromere is shorter for *a* than for *c* and, accordingly, that *a* is closer to the centromere of chromosome 9 than *c*. In the same sort of cross, it is also possible to assess the distance between the centromere and the most proximal marker (*a* in this case). These distances, however, are relatively unreliable

translocations. The centric fusion can be identified in a karyotype, and its presence or absence is used as phenotypic information that can be integrated into a mapping experiment with the phenotype of other mutations on the same chromosome. When, for example, a mutant allele segregating in a cross has a tendency to stay associated with the metacentric chromosome in the same parental configuration (i.e., in coupling or repulsion), this means that the centromere is very likely in the close vicinity of the locus for the mutant allele in question (Fig. 4.2).

In the sort of cross we just described, the distance between the centromere and the nearest genetic locus could in principle be estimated by counting the recombinant/non-recombinant genotypes. Unfortunately, it has been demonstrated that in crosses where this sort of translocation segregates, the distances to the centromere are frequently underestimated due to structural differences that interfere with meiotic pairing (Davisson and Akeson 1993). The positioning of the centromeres was rapidly established. It took a little longer to compute the distance between the centromere proper and the first (proximal) locus.

4.3 Genetic Markers

Any gene or short DNA sequence whose chromosomal location is precisely known and whose structure exhibits variations among the individuals of the same strain or species can be considered a *genetic marker*. The ancestral mutation *albino* is a good example to explain this concept. The mutation has been cloned and found to be the consequence of a G-to-C transversion at nucleotide 308 of the gene encoding tyrosinase (*Tyr*-Chr 7) (Yokoyama et al. 1990). The nucleotide substitution results in the replacement of a cysteine residue by a serine (a missense mutation), making the product of the *Tyr* gene unable to cooperate in the production of the pigment melanin, thus explaining why the affected mice are albino. Looking in more detail at the sequence of the Tyr^{c} gene, it was also noticed that the nucleotide substitution introduced a novel *Dde*I restriction site. In these conditions, the coat color (albino), the G-to-C transversion and the *Dde*I restriction site are all genetic markers whose polymorphism can be used to track the same *Tyr* allele (i.e., the same DNA sequence of Chr 7), generation after generation. The coat color of the mouse is a *phenotypic marker*, while the restriction site and the nucleotide substitution are *molecular markers* (DNA markers, in this case) of exactly the same locus.

As we explained at the beginning of this chapter, when mouse genetics began, the markers were exclusively phenotypic, i.e., were identified as visible changes such as an alteration of the coat color or fur texture, or skeletal anomalies, or abnormal behavior, etc. With the progress of biochemistry, molecular markers were developed that proved to be more abundant and easier to characterize. In addition, most of these molecular markers are co-dominantly expressed, rather than dominant/recessive like phenotypic markers. Nowadays, most of the genetic markers are scored at the DNA level, and are abundant and easy to characterize

with highly reliable techniques. These DNA markers represent small structural changes at the DNA level, distributed over the whole genome, including the coding and non-coding regions. They have allowed the very rapid expansion of genetic maps in all mammalian species.

4.3.1 Markers Scored by Examination of the External Phenotype

The mouse genetic (or linkage) map originated from the observation that the albino (*Tyr*^{c}) locus was linked to the pink-eyed dilution locus (*p*—now *Oca2*^{p}). After this initial observation, the linkage map expanded for over 60 years as a consequence of the continuous discovery of new mutant alleles, dispersed at different loci throughout the mouse genome, and meticulously mapped, one at a time.[18, 19]

All these mutant alleles, with an obvious phenotypic effect, could be used as genetic markers, for example for the mapping of a new heritable trait. Unfortunately, they have two major drawbacks. The first and most important is that they often impair the viability and/or the fertility of the affected animals in one sex or in both, and for this reason it is extremely difficult to set up crosses (especially testcrosses) involving more than three (maximum four) markers of this kind. This would make the chromosomal assignment of a new trait time-consuming and expensive, not taking into account the sacrifice of many animal lives. Second, even if all these new mutant phenotypes, by their abundance, were revealed to be important for establishing the frame (or scaffolding) of the genetic map, they would nevertheless remain insufficient for developing the high-density map that would be indispensable for the analysis of the whole genome. For these reasons, genetic markers detectable by examination of the external phenotype are rarely used.

---

[18] Accumulation of these new mutant alleles was, in part, a direct consequence of the use of the mouse as a model organism for the evaluation of the effects of radiation or of chemical mutagens on the genome, and an indirect consequence of inbreeding as a mating system. Inbreeding has no effect on the mutation rate, but it increases the chance that individuals will be homozygous for recessive mutations, which, accordingly, are more easily identifiable.

[19] During the years 1965–1975, when the genetic map of the mouse was expanding, researchers at Harwell MRC and The Jackson Laboratory were using the so-called *linkage testing (or linkage tester) stocks*. These stocks were homozygous for up to seven carefully selected recessive, fully viable, and fully penetrant coat color markers mapping to different chromosomes. The phenotype of each of these markers could be detected independently, with no interference from the other markers. One of these stocks was the famous PT stock, which was extensively used at Oak Ridge by W.L. & L.B. Russell for estimating the rate of induced mutations either with chemicals or radiation. The PT stock was homozygous for seven markers on five chromosomes: *a/a; b/b; c*^{ch}*-p/c*^{ch}*-p; d-se/d-se; wa1/wa1*.

4.3.2 Electrophoretic Variant of Enzymatic Proteins

From the late 1960s onwards, the development of gel electrophoresis and the concomitant discovery of techniques for staining the product(s) of enzymatic reactions allowed the identification of polymorphisms resulting from discrete variations in the electrical charge of enzymatic proteins. These types of molecular markers, referred to as electromorphs or electrophoretic variants, were found to be very convenient for the purpose of mapping because they have three advantages over the phenotypic markers described above: (i) they are co-dominantly expressed and can thus be independently typed in heterozygotes; (ii) they are in general compatible with a normal function of the enzyme and accordingly do not impair the viability or fertility of the animals; (iii) they are relatively abundant, probably because they are selectively neutral, and therefore allow a wide coverage of the genome. Well over 100 markers of this kind are available and new variants are regularly discovered, particularly in wild mice. These biochemical markers permitted a rapid expansion of the mouse linkage map in the mid-1970s (Bonhomme and Selander 1978). Unfortunately, they have the drawback that their characterization requires relatively sophisticated techniques, making large-scale linkage experiments based on this approach quite expensive to run. Nowadays, these markers have been virtually abandoned (Fig. 4.4a).

4.3.3 Plasmatic Proteins and Cell Surface Antigens

Many plasmatic proteins (albumin, globulins, etc.) or tissue-specific proteins (seminal vesicle proteins, lens proteins, urinary proteins, etc.) and cell surface antigens have also been used as genetic markers, with the same advantages and drawbacks as the electrophoretic variants. They can still be used, but they do not exhibit any obvious advantages over the molecular (DNA) markers geneticists now have at their disposition.

4.3.4 Polymorphisms Detected at the DNA Level

The progress in DNA technology accomplished in the early 1980s has permitted the development of two new kinds of genetic markers: (i) those that are generated by restriction endonucleases and detected by Southern blotting: the so-called *restriction fragment length polymorphisms* (RFLPs), and (ii) those that are detected by DNA amplification using the polymerase chain reaction (PCR).

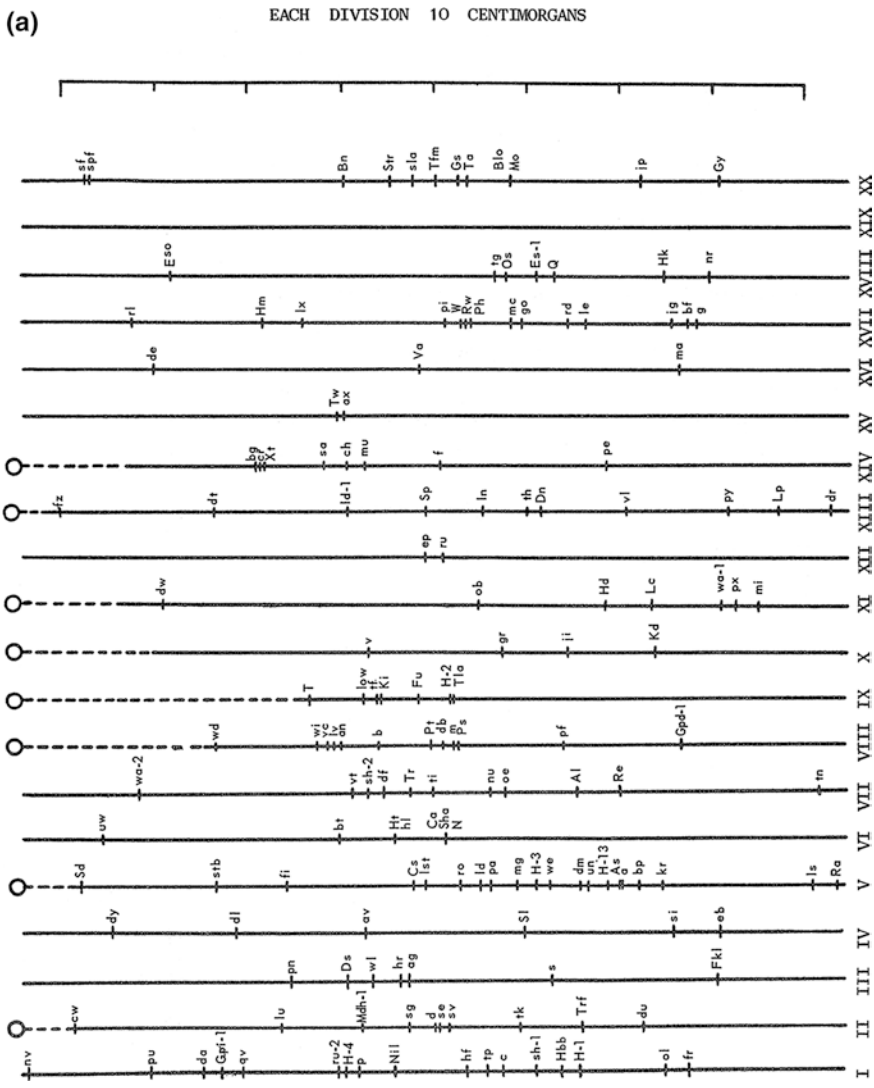


Fig. 4.3 **a** *Linkage map of the mouse as it appeared in 1971*. This map was compiled and kept up-to-date by Dr. Margaret Green from The Jackson Laboratory, Bar Harbor, Maine, USA. The only LG that was assigned to a chromosome was LG XX (assigned to Chr X due to sex linkage). **b** *Genetic map of the mouse as it appeared in the Mouse News Letter in 1975*. This map was also compiled by Dr. Margaret Green. The genes of LG XVI were incorrectly assigned to Chr 12 and should have been assigned to Chr 3. All other assignments were correct and the centromere was also correctly positioned

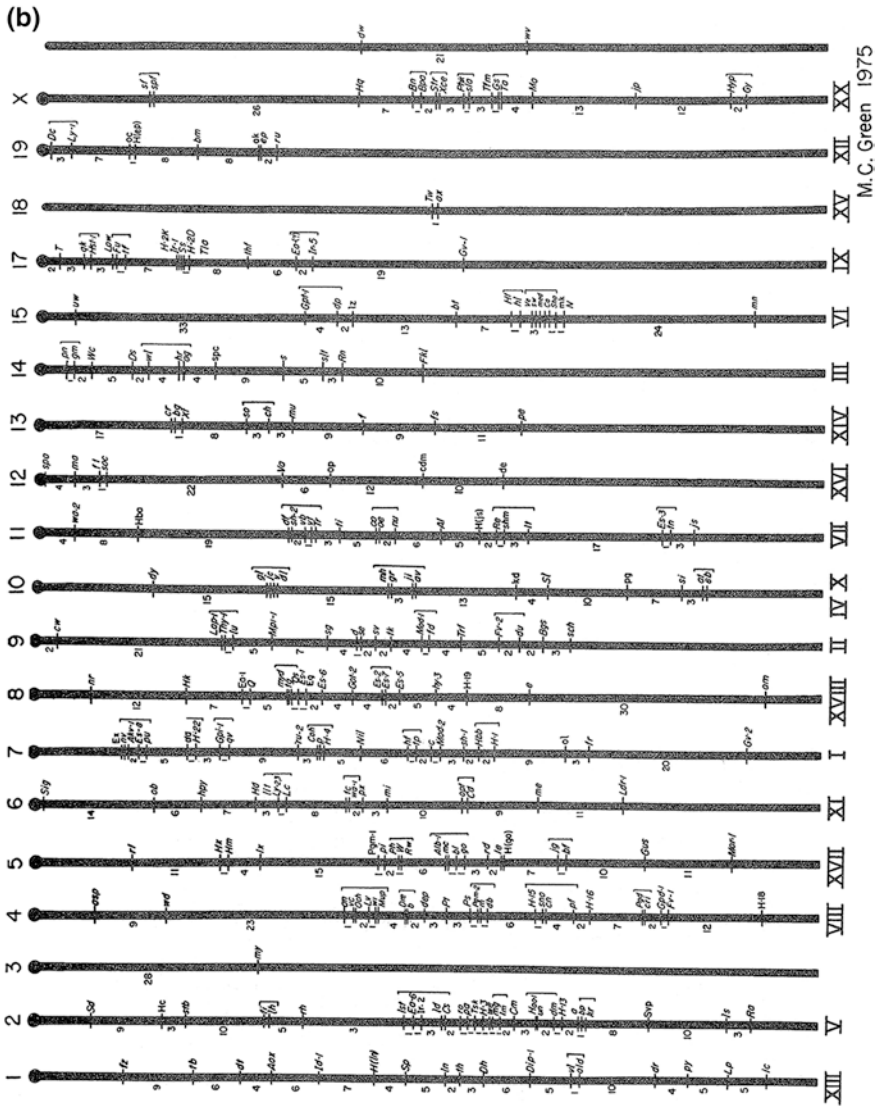


Fig. 4.3 (continued)

4.3.4.1 Restriction Fragment Length Polymorphisms

In 1982, Botstein et al. (1980) reported that the restriction fragments generated by digestion of DNA samples from different individuals, with one of the various restriction endonuclease, often exhibited size polymorphisms when observed with the technique of Southern blotting. These restriction fragment length

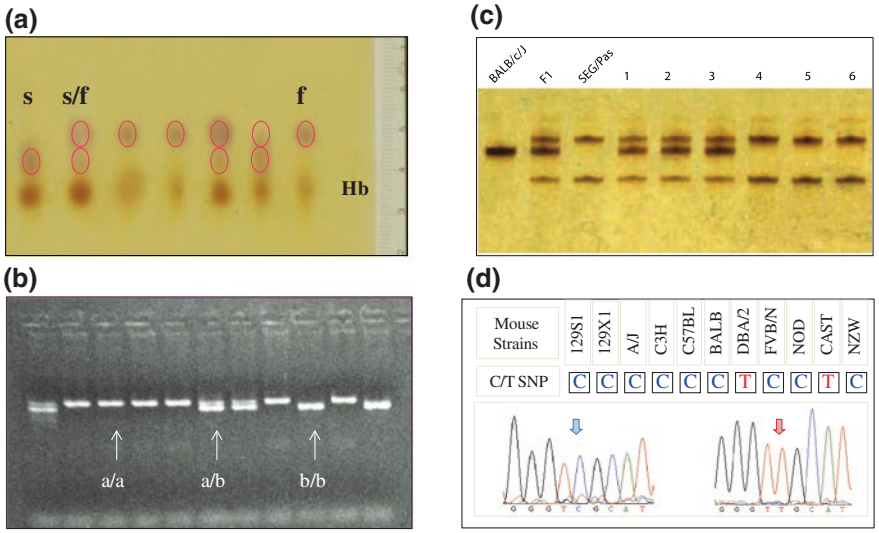


Fig. 4.4 *A sample of co-dominant molecular markers*. The picture represents four kinds of genetic markers, selected from those that have been (or still are) the most frequently used. **a** represents a gel discriminating two electrophoretic variants of the enzyme adenosine deaminase (*Ada*-Chr 2). The slow variant (*s*) exists in mice of the *Mus spretus* species (strain SEG/Pas). The fast variant (*f*) is common in classical laboratory strains. The F1 (s/f) between SEG and the laboratory strain synthesizes the two forms. **b** represents the size polymorphism observed with microsatellite *D19Mit1* and DNA samples from two laboratory inbred strains. Microsatellites are extremely abundant in mammalian genomes. They consist of a tandemly repeated, short-sized motif, usually 2–6 bp long. Genotyping of this class of marker is achieved by PCR amplification using primers designed from the flanking sequences, while the size differences of the amplification products are assessed in agarose or polyacrylamide gels. Heterozygous genotypes are clearly visible. **c** represents a single strand conformation polymorphism or *SSCP* assay. Once denatured by heat (~90 °C), single-stranded DNAs exhibit a three-dimensional folding that is influenced by their sequence. The spatial conformation of the single-stranded DNA determines its ability to move in the gel (acrylamide). SSCP is an extremely sensitive technique but is now being supplanted by sequencing techniques on account of efficiency and accuracy. **d** represents a single nucleotide polymorphism (*SNP*). This type of polymorphism is extremely abundant

polymorphisms or RFLPs also behave as co-dominant markers, and accordingly have been extensively used for linkage analysis. Restriction endonucleases are extremely abundant, and every event that suppresses or creates a restriction site in the DNA sequence can be regarded as a creator of a new genetic marker. RFLPs can be detected independently of their position in the DNA chain, be it within or outside the coding sequences. The RFLP method was a true breakthrough because it allowed mapping of virtually any locus encoding for a protein once its DNA sequence was known. For this reason, it permitted the very rapid expansion of linkage maps in the 1990s (Minty et al. 1983). Nowadays, the detection of RFLPs by Southern blotting has been virtually abandoned, but checking for the presence/absence of a restriction site in a PCR amplification product may still be used.

4.3.4.2 Markers Detectable by PCR

Several other techniques have been reported that are also based on the analysis of structural variations at the DNA level. Among the most interesting are those taking advantage of PCR, because they require very small amounts of template DNA and because the genotyping can be completed within only a few hours at a relatively low cost. The most popular of these techniques consists of the amplification of short sequences (usually less than 300 bp) whose polymorphisms are either in size [*simple sequence length polymorphisms* (SSLPs or microsatellites)] or in the sequence itself [*single strand conformation polymorphisms* (SSCPs), *denaturing gradient gel electrophoresis* (DGGE)]. Many other DNA markers have been described and used in the mouse and found to be interesting, either because of their simplicity or because they provided geneticists with an almost unlimited number of markers at very low cost, for example the RAPDs (Serikawa et al. 1992). Unfortunately, these markers had the drawback of not being repeatable from one experiment to the next, and ultimately they did not present definitive advantages over the microsatellites. For this reason they were only used briefly.

Microsatellites

Microsatellites [also known as *simple sequence repeats* (SSRs), *short tandem repeats* (STRs), or *simple sequence length polymorphisms* (SSLPs)], correspond to the repetition in tandem of relatively short motifs, usually 2–6 bp long.[20] They are co-dominant markers. They are commonly found in the mammalian genomes and have been extensively used for the purpose of mapping in a variety of species. Their origin is debated but it makes sense to think that they result from errors occurring during DNA replication, a sort of stuttering or slippage of the DNA polymerase, or from unequal recombination events. The size polymorphisms are generally assessed by analysis of the migration of the amplification products after electrophoresis in agarose or polyacrylamide gels. The primers used for amplification are designed to target the unique sequences flanking the repeats (Love et al. 1990; Hearne et al. 1991; Montagutelli et al. 1991) (Fig. 4.4b).

These microsatellites represent almost ideal molecular markers because: (i) they are highly polymorphic among strains; (ii) they are usually found in the non-coding regions and the polymorphisms therefore are less likely to alter phenotype of the mouse; (iii) they are abundant (in the range of 10^5 copies in the mouse genome, at least for CA repeats); and (iv) they are relatively stable generation after generation.[21]

---

[20] Repeated units such as T, CA, CT, and CAG are among the most common.

[21] In fact, the microsatellites are more mutable than most other molecular markers previously described, but their mutation, in general, generates a novel allele that is not identical to any of the parental alleles. In these conditions the structural instability does not result in a confusion. It may, however, be a problem when microsatellites are used for the genetic monitoring of inbred strains (see Chap. 9).

Nowadays, primer sequences for thousands of SSLP markers are deposited in public databases, and sometimes even commercially available. These markers can be used either for the mapping of new polymorphic loci on the mouse linkage map or for the identification of a DNA clone in a library.

When using these markers for the genotyping of DNA samples from inter-specific or inter-subspecific crosses, one must be careful and make sure that the selected primers prime amplification equally well with DNA of the two parental strains or species. Not taking this warning into account may lead to confounding results, with heterozygous genotypes being incorrectly classified as homozygous. This warning, of course, applies to all PCR assays performed on DNAs of different species and is not specific to the SSLPs, even though it is frequently observed in this particular case.

Single Strand Conformation Polymorphisms

Sequence polymorphisms of PCR products can also be evaluated by comparing their electrophoretic mobility in a polyacrylamide gel, after denaturation into a single-stranded form (SSCP) or by measuring their migration abilities within a gel containing a gradient of denaturing compounds such as urea or formamide (DGGE). In these two cases, the method is so sensitive that a difference of only a single base-pair in the sequence of any DNA molecule 80–250 bp long is generally detectable (Fig. 4.4c).

Single Nucleotide Polymorphisms

After the sequencing of the mouse genome, a new kind of genetic marker has been developed: the *single nucleotide polymorphisms* (SNPs—see Chap. 5 for a full description). These polymorphisms are single base-pair changes occurring throughout the genome in all sorts of sequences (coding and non-coding). They are extremely abundant and stable, and SNP genotyping is available on different platforms including real-time PCR (TaqMan®), DNA microarrays, and competitive allele-specific PCR coupled with fluorescence resonance energy transfer technology (Livak 1999; Nijman et al. 2008; Yang et al. 2009). However, sequencing short DNA stretches (e.g., using pyrosequencing) in search of SNPs is still an alternative approach to small-scale projects. SNPs are generally bi-allelic, meaning that there are generally no more than two alleles across the different strains or species of the genus *Mus*. Petkov and coworkers from The Jackson Laboratory have described the allelic distribution of 235 SNPs in 48 mouse strains and selected a panel of 28 such SNPs, enough to characterize hundreds of strains (Petkov et al. 2004a). The same laboratory developed a new set of 1,638 informative SNPs selected from the publicly available databases and tested 102 inbred strains (Petkov et al. 2004b). For those interested in the allele distribution of SNPs in different inbred strains, the Mouse Phenome Database (MGD) presents

the most comprehensive collection of SNPs, with more than 8 million unique loci and numerous inbred strains genotyped. In short, one can say that nowadays, with microsatellites and SNPs, genetic markers are very abundant and relatively cheap to characterize (Fig. 4.4d).

4.4 High-Resolution, High-Density Genetic Maps

When discussing the strategies used for evaluating the genetic distances between genes or markers, we explained that the precision of these distances depended upon the number of mice scored. Every mouse in a backcross progeny represents a certain number of independent recombination events, randomly distributed over the whole genome, and the greater the number of mice scored, the more information is collected. Theoretically, and not taking statistical variations into account, genotyping 100 offspring of a backcross progeny should be sufficient to find, on the average, one mouse whose genome is recombinant between two genes or markers distant by 1 cM (or ~1.7 Mb of DNA). If we increase the number of backcross progeny, for example 10 times, and score 1,000 mice instead of only 100, then the theoretical resolution of our mapping would rise up to the milliMorgan level (equivalent to 170 kb of DNA). In this case our map would be considered a *high-resolution* genetic (or meiotic) map, meaning that it is very precise. Such a map would be very useful for anchoring overlapping cloned DNAs covering a chromosomal region. However, and unless we are able to precisely localize all the thousands of cryptic recombination events, such a map will not be very helpful. In other words, by genotyping 1,000 backcross mice we would be able to compute with great precision the distances between the few markers that are polymorphic in the backcross, but nothing else. In contrast, if we collect a few hundred mice from a backcross set up between two remotely related inbred strains belonging to two sub-species of the *Mus* genus (*Mus m. domesticus* × *Mus m. castaneus*, for example), we may then establish a high-resolution and *high-density* map because, as explained earlier, thousands of markers of all sorts would be polymorphic in such a cross.[22]

Several high-resolution/high-density maps were developed in Europe and in the USA in the late 1990s (Dietrich et al. 1994; Rhodes et al. 1998). Among the most important resources of this kind is the *European Collaborative Interspecific Backcross* (the EUCIB resource), which was established from a collection of 982 DNA samples prepared from the progeny of two large backcrosses involving mice of *Mus spretus* species and of the C57BL/6 inbred strains. This resource, which incorporates 3,368 microsatellite markers distributed among 2,302 genetically

[22] Around 70 % of SSLPs (microsatellites) or SNPs have been found to be polymorphic between any two strains derived from progenitors of independent (wild) origins of the same *Mus* genus. Altogether, this means that around 30,000 SNPs or SSLPs could (potentially) be used for the purpose of mapping in a cross involving two inbred strains derived from two different subspecies.

separated bins, with 1.46 markers per bin on average, allowed mapping any DNA with a genetic resolution of 0.3 cM at the 95 % confidence level (approximately 600 kb in the mouse genome).[23, 24]

High-resolution/high-density maps have been used for assembling the cloned DNAs in a physical map, and accordingly have logically contributed to the establishment of the mouse genome sequence.[25]

4.5 Somatic Cell Hybrids and Radiation Hybrids as Tools for Gene Mapping

In the period spanning 1975–1985, geneticists used hamster/mouse *somatic cell hybrids* to assign mouse genes to a particular chromosome. These cell hybrids were obtained by merging in vitro somatic cells (fibroblasts) of the two species, the fusion being triggered by small amounts of polyethylene glycol (PEG) (Pontecorvo 1976). These cell hybrids were unstable, progressively and randomly losing the chromosomes of the mouse. After a few generations of in vitro culture, the situation stabilized and some of the clones remained stable, with a full set of intact hamster chromosomes plus some extra mouse chromosomes or fragments of mouse chromosome in their karyotype. The set of mouse chromosomes retained in the individual clones was identified by one of the techniques described in Chap. 3, and was in most instances unique for each clone. In these conditions, the cell hybrids could be used to perform the rapid chromosomal assignment of mouse-specific markers accessible in vitro (protein or a DNA polymorphism), by matching the results of typing for the different clones, expressed in terms of presence (1 or +) or absence (0 or –), with the karyotype of the same clones and the presence or absence of a specific mouse chromosome. This method of somatic cell hybrids was relatively global, and provided only a chromosomal assignment.

The method was greatly improved in the early 1990s by replacing the mouse cells with cells of the same species irradiated with a sub-lethal dose of X- or γ-rays. In these conditions it was fragments of mouse chromosomes

[23] A bin is a group of syntenic genetic markers that have not been separated (ordered) by meiotic recombination in a given cross.

[24] The mapping of these microsatellites has been achieved in two successive steps. First, all the 982 DNA samples of the backcross progeny were initially typed for 78 primary anchor loci spanning the entire genome, with 3–6 anchors per chromosome. In a second step, only the DNA samples demonstrated to be recombinant in one or the other of these intervals tagged by the 78 primary anchor loci were typed for the greatest possible number of markers located (or presumed to be located) in these regions.

[25] In fact, and as we shall see in Chap. 5, the mouse genome has been sequenced by using a global strategy known as *whole-genome sequencing* or WGS. However, the physical map of the mouse genome, established by anchoring a variety of DNA clones on the genetic map, has been helpful in many experiments of positional cloning, and is still used for the analysis of quantitative traits.

instead of complete chromosomes that were retained in the genome of the non-irradiated hamster cells. The size of these fragments was inversely correlated to the dose of irradiation. In the case of the mouse–hamster T31 panel, the panel of radiation hybrids (RH) that was mostly used during this period, the dose of irradiation was 30 grays, generating fragments measuring 10 Mb on average. In these conditions, when several hybrid clones were discovered to exhibit the same pattern of presence/absence for a specific marker, this was suggestive of linkage to the same fragment of chromosome inserted somewhere in a hamster chromosome. Compared with the other mapping strategies, the use of RH for high-resolution mapping had some undisputable advantages. The first is that it did not depend upon meiotic recombination and accordingly did not require any cross between animals. Another advantage is that the method could be used even in the cases where the cloned DNA to be mapped did not exhibit any polymorphism. Finally, the results gathered from one experiment contributed to the enrichment of the database associated with the panel of RH cells used.

The method has been extremely helpful for defining gene order and distances in the range of 4–8 Mb (i.e., equivalent to ~2–5 cM), especially in the later phases of the development of the mouse genetic map, for the establishment of the high-resolution/high-density consensus maps. At the end of 2001, the map established with the help of the T31 RH panel contained up to 11,109 markers, positioned relative to a reference map containing 2,280 genetic markers. It included 3,658 genes homologous to the human genome sequence. Nowadays, given that the mouse genome is completely sequenced, the method no longer has an application. Indeed, no map can have a better resolution than the sequence itself. For more information concerning this non-sexual mapping strategy, the following references are recommended: Cox et al. 1990; McCarthy et al. 1997; Flaherty and Herron 1998; Hudson et al. 2001.

4.6 Recombinant Inbred and Recombinant Congenic Strains

The nature of recombinant inbred strains (RIS) and their importance in mouse genetics are described in detail in Chap. 9. Here we will only discuss the interest of these strains as a tool for gene mapping.

RIS are derived from two unrelated parental inbred strains by systematically intercrossing (brothers × sisters) the successive offspring of pairs of F1s for at least 20 generations and often many more (Bailey 1971; Taylor 1978; Williams et al. 2001). Figure 4.5 illustrates the genetic structure of such strains. RIS derived from the same parental strains go by sets, or panels. At the present time, the BXD panel, derived from the inbred strains C57BL/6 and DBA/2, is the largest panel with ~90 strains available for research, but several other smaller-sized panels have also been developed.

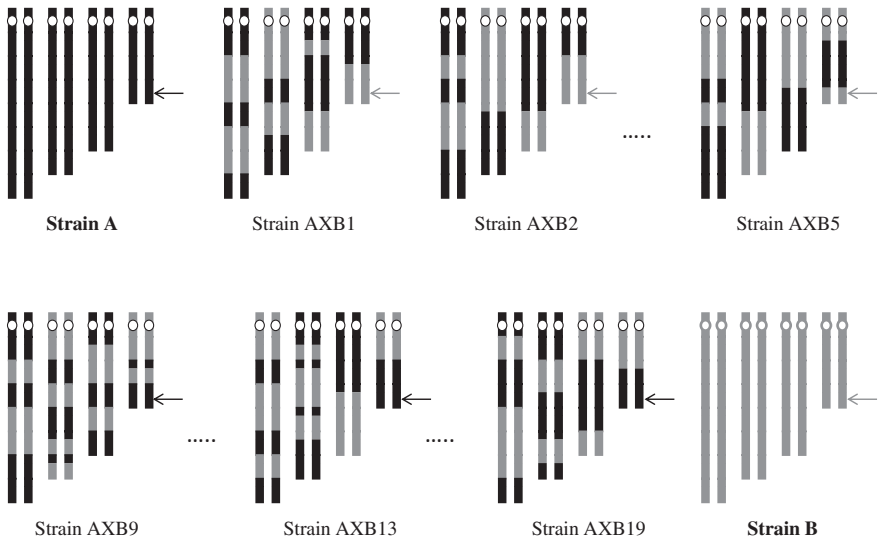


Fig. 4.5 *The genetic structure of recombinant inbred strains*. RIS are derived from two parental inbred strains (here Strain A and Strain B) and propagated by strict inbreeding of several pairs of inter-strain (A × B) F1. After 20 or more generations, these strains are totally inbred and each chromosome of their genome is a patchwork of the two parental components. A given locus has either the A or the B allele, as exemplified by the *arrow*. These RIS have been useful for the purpose of mapping molecular markers, and are still extremely helpful for the analysis of the genetic determinism of the quantitative traits (QTLs) that are different in strains A and B

Just like the parental strains they are derived from, RIS are also inbred, meaning that within a given strain all individuals are genetically identical. However, each of them has a unique combination of the parental alleles. For example, any strain of the BXD panel carries either the C57BL/6J (B) or the DBA/2J (D) allele (in homozygosity) at any given locus of its genome, in a 50:50 ratio. By typing all of these allelic forms at every locus and for each of the strains, one establishes what geneticists call a *strain distribution pattern* (SDP). The SDP is a permanent source of information that is progressively implemented after every new genotyping. It is also a basic characteristic of the panel of RIS.

RIS have proven to be excellent tools for mapping mouse genes, and nowadays the strategy is in expansion for the mapping of quantitative traits (Fig. 4.5). The reasons for this success are two-fold. (i) Each strain in a panel of RIS represents a collection of individuals with identical genomes that can be bred in unlimited numbers. These strains remain stable generation after generation, with the only exception of possible rare new mutations. Samples of animals of a given strain can then be phenotyped and genotyped very accurately, for all sorts of characteristics including quantitative traits, molecular markers, etc. A great advantage of RIS is that large homogeneous samples of animals can be prepared for genotyping. Another advantage is that recurrent phenotypings are also possible. (ii) Having origins in two unrelated progenitor strains, the genome of each strain looks like

a patchwork made up of chromosomal fragments derived, randomly, from one or the other of the two progenitor strains. In these conditions it is clear that the genes linked on a given chromosome have a tendency to remain associated in the same parental configuration through the successive generations of inbreeding, except when a crossing-over splits the association. While inbreeding progresses, crossing-over events occur at every generation, in each sexual partner, which modify the genotype of the strain as long as the chromosomal segment stays heterozygous. The chromosomes are thus progressively sliced into smaller-sized segments. In the end, the genetic structure of the strain stabilizes and each individual strain of the panel is homozygous for, on average, 50 % of the alleles coming from one parental strain and 50 % of the alleles coming from the other parental strain, but each strain has a unique patchwork.

As we already mentioned, the phenotyping and genotyping information collected in a panel of RIS can be used additively. The results collected for markers *A* and *B*, for example, can be used for the mapping of markers *C*, *D*, etc..

The most popular panels are BXD (inbred parents C57BL/6J and DBA/2J), AXB-BXA (inbred parents C57BL/6J and A/J), AKXD (inbred parents AKR/J and DBA/2J), and AKXL (inbred parents AKR/J and C57L/J). Other panels are also available with only a few strains (CXB, BXH, etc.), but they are also useful.

When a new marker or a new phenotype exhibits differences between the parental strains, a novel strain distribution pattern (SDP) can be established and matched to the SDPs already stored in the databases, in general with the help of a computer program. The position of the new marker or phenotype and its linkage with flanking markers is then calculated. The principle for positioning a new marker or phenotype is to generate the smallest possible SDP discordances with the flanking markers. As we already said above, when the marker density is very high in a chromosomal region, ordering of the different loci is sometimes impossible and, in this case, the group of unordered, syntenic markers represents a bin (Fig. 4.6).

If the panel of RIS is large enough, the genetic distances can be calculated based on standard formulas published in the genetic literature (Bailey 1971; Taylor 1978; Silver 1985). For example, if the number of discordant strains for a pair of adjacent (linked) loci is *i* and the total number of RIS is *N*, one has:

$$R = i/N$$

Once *R* is established from the analysis of the SDPs, the recombination frequency *r* can be estimated from the formula:

$$r = R/(4-6R)$$

Tables are available that provide an estimation of the percentage recombination *r* (equivalent to the genetic distance) between two loci, and the upper and lower 95 % and 99 % confidence limits for *r*, when $6 \leq N \leq 45$) (Silver 1985).

For example, if 5 RIS, out of a total of 23, are found to be discordant in their SDP for markers *A* and *B*, two genes on the same chromosome, then $i = 5$, $N = 23$, the formula returns a recombination fraction (*r*) of 0.0806 (=8.06 %), and

	2	5	6	8	9	11	12	13	14	15	16	18	19	20	21	22
Locus 1	B x	B	B	D	B	B	B x	B	B	D	B	B	D	B	B	D
Locus 2	D	B	B	D	B	B x	D	B	B	D	B	B	D x	B	B	D
Locus 3	D	B x	B	D	B	D	D	B	B	D	B x	B	B	B	B x	D
Locus 4	D	D	B x	D	B x	D	D	B	B	D	D	B	B	B	D	D x
Locus 5	D	D	D x	B	D	D	D	B	B	D x	B	B	B x	D	D	B
Locus 6	D	D	B	D	D x	D x	B	B x	B	B	B x	D	D	D	B	
Locus 7	D x	D	D	B	D x	B	B	D	B	D	D	B x	D	D	B	
Locus 8	B	D x	D	B	B	B	B x	D	B	D	D	D	D	D	B	
Locus 9	B	B	D	B	B	B x	B	D	D	B	D	D	D	D	B	
Locus 10	B	B	D	B	B x	D	B	D	D	B x	D	D	B	D x	B x	
Locus 11	B	B	D x	B x	D	D	B	D	D	D	D x	D x	B	B	D	
Locus 12	B x	B	B	D	D	D	B	D	D	D	B	D x	B	B	D	
Locus 13	D	B	B	D	D	D	B	D	D	D	B	B	B	B	D	
Locus X	**D**	**D**	**D**	**B**	**D**	**B**	**D**	**B**	**D**	**B**	**D**	**B**	**D**	**D**	**B**	

	2	5	6	8	9	11	12	13	14	15	16	18	19	20	21	22
Locus 1	B x	B	B	D	B	B	B x	B	B	D	B	B	D	B	B	D
Locus 2	D	B	B	D	B	B x	D	B	B	D	B	B	D x	B	B	D
Locus 3	D	B x	B	D	B	D	D	B	B	D	B x	B	B	B	B x	D
Locus 4	D	D	B x	D	B x	D	D	B	B	D	D	B	B	B	D	D x
Locus 5	D	D	D	D x	D	D	D	B	B	D x	B	B	B x	D	D	B
Locus 6	D	D	D	B	D	D x	D	B	B x	B	D	B	D	D	B	
Locus X	**D**	**D**	**D**	**B**	**D**	**B**	**D x**	**B**	**D**	**B**	**D**	**B x**	**B**	**D**	**D**	**B**
Locus 7	D x	D	D	B	D x	B	B	D	B	D	D	B x	D	D	B	
Locus 8	B	D x	D	B	B	B	B x	D	B	D	D	D	D	D	B	
Locus 9	B	B	D	B	B	B x	B	D	D	B	D	D	D	D	B	
Locus 10	B	B	D	B	B x	D	B	D	D	B x	D	D	B	D x	B x	
Locus 11	B	B	D x	B x	D	D	B	D	D	D	D x	D x	B	B	D	
Locus 12	B x	B	B	D	D	D	B	D	D	D	B	D x	B	B	D	
Locus 13	D	B	B	D	D	D	B	D	D	D	B	B	B	B	D	

Fig. 4.6 *Mapping a trait with RIS*. This table represents the (theoretical) results of the genotyping (or phenotyping) for each strain of a panel of 16 RIS (*top row*). This panel is commonly designated as the Strain Distribution Pattern or SDP. All strains are homozygous for one or the other allelic forms present in the parental strains (in this case, B for C57BL/6 or D for DBA/2). This SDP is permanent information that can be easily found in the public databases. When the parental strains are discovered to differ for a particular phenotype (or genotype), the panel can then be used for mapping this new characteristic. Each strain is typed as B (identical to parent C57BL/6) or D (identical to parent DBA/2), and the new SDP (*Locus X*) is plotted with the existing data looking for the highest possible concordances. This is generally achieved very rapidly by using simple, publicly available software. In the case shown, the best position for *Locus X* (or phenotype *X*) is between locus 6 and 7. Nowhere else could the SDP be more similar to the neighboring loci (two discordances with locus 6 and two with locus 7)

the table gives 2.1 and 31.7, respectively, as lower and upper confidence limits at the 5 % risk. With the same number of discordant strains ($i = 5$) but a larger number of RIS ($N = 44$, for example), the same table would give 0.0342 (= 3.4 %) for r, with 1.0 and 9.7 as lower and upper confidence limits at the same 5 % risk.

The strategy making use of the RIS "expands" the map over short distances, and is more efficient than a backcross population for estimating recombination when map distances are relatively small (<12.5 cM).[26] On the other hand, cross–intercross or cross–backcross protocols are more appropriate for the detection of linkages over distances in the range of 20–30 cM (Silver and Buckler 1986).

RIS have proven extremely helpful for the rapid regional assignment of microsatellites on a given chromosome when these markers were cloned by the thousands for the establishment of high-density genetic maps (Dietrich et al. 1994). They have also been used for the mapping of chromosomal regions (QTLs) involved in the genetic determinism of several behavioral characteristics (for example, taster/non-taster for chemical compounds, alcohol intake, etc.) or for the mapping of some immunological responses, or susceptibility to pathogens. They will very likely be of great help in many other experiments where the phenotype is measured on a group of animals rather than on individuals (Zou et al. 2005).

[26] This is sometimes referred to as a "magnifying glass effect.".

Recombinant congenic strains (RCS) are similar to RIS in their genomic structure, except that the proportion of the parental alleles in a given panel is not 50:50 but 75:25 or even 87.5:12.5, depending on the panel (Demant and Hart 1986). These RCS are produced by crossing mice of the first or second backcross generation to one of the parental inbred strains (the background strain), followed by strict inbreeding. As we will explain in Chap. 10, RCS are helpful for identifying genes or QTLs, especially when the latter are numerous. RCS with a small percentage of introgressed genome in a background strain have a greater power of resolution, and their use increases the likelihood of no or only a single locus (or QTL) governing the phenotype being isolated in a given strain. For example, RCS have been very helpful for unraveling the genetic determinism of colon cancer in the mouse (Demant 2003). Interspecific recombinant congenic strains (IRCS) have also been developed from the parental strains C57BL/6JPas and SEG/Pas (*Mus spretus*) (Burgio et al. 2007). This set of strains has permitted the analysis of the genetic architecture of some anatomical traits (Burgio et al. 2012).

4.7 Establishing Consensus Maps

In the previous sections of this chapter, we explained that the genetic localization of genes or cloned DNAs could be achieved by different methods. Two of these methods are based on meiotic recombination (linkage maps and RIS), while the third is based on the analysis of the retention of mouse chromosome fragments of various sizes in hybrid cells. These methods were extensively used by mouse geneticists at the end of the twentieth century and have contributed to the construction of rich and dense maps. However, problems arose when the point came to merge the mapping data collected through one of the three methods into one and the same *consensus map* (Fig. 4.7).

When embarking on such a project, a few points must be taken into account:

- The construction of a consensus map based on independent primary maps is possible only when a number of markers (designated *anchor* markers) are common to all primary maps.
- The distances computed between loci are not the same for meiotic and for RH maps, but the distances computed by RH mapping are closer to the physical distances than those provided by linkage maps.
- The genetic distances computed from meiotic recombination depend upon the crosses. As we already said, meiotic distances computed from male meiosis are in general not the same as distances computed from female meiosis. We must also mention that crosses involving progenitors from different strains, or from different subspecies, sometimes result in distortions in the genetic distances. This is generally the consequence of small differences in the chromosomal structure (inversions, deletions, etc.) and must be taken into account (Paigen and Petkov 2010).

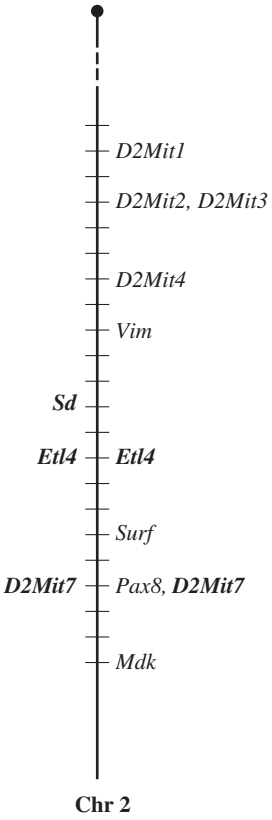


Fig. 4.7 *Consensus map*. The observation of degenerative changes of the notochord in mice homozygous for Danforth's short tail dominant mutation (*Sd*) at day 10 of development, and the concomitant localization of this mutation in the proximal region of Chr 2 were strong arguments for making the gene encoding for paired box 8 (*Pax8*) a likely candidate for *Sd*. Merging the data of two genetic maps, one involving the mutant allele *Sd* and the molecular markers *Etl4* and *D2Mit7* (*left*), the other involving several molecular markers including *Pax8* (*right*) into a consensus map, ruled out this hypothesis because *Sd* was found to be proximal to *Etl4* while *Pax8* is distal to the same marker. This consensus map could be established only because two markers (*Etl4* and *D2Mit7*) were common to the primary maps (Redrawn from Koseki et al. 1993)

- The maps established by analysis of the SDP of RIS have gaps. These gaps are inherent to the origin of the set of RIS and result from an absence of genetic polymorphism between the parental strains in some chromosomal regions.
- The mapping information collected from one set of RIS (order and distances) can be merged with the data collected from another set, provided that this refers to the same markers or genes.
- Data relative to gene order in the syntenic regions of other mammalian species are important to consider, especially when the species are closely related.

However, they must be used only as an indication. The genetic distances cannot be compared because the recombination frequencies are species-specific.

In the years 1991–1999, mouse geneticists established consensus maps by gathering the largest possible amount of mapping data from the literature. These maps were mostly molecular maps excluding most of the mutant phenotypes, unless the mutant alleles were cloned. These consensus maps, compiled by chromosome committees, have been published in several successive special issues of the journal *Mammalian Genome* between 1991 and 1999 (Fig. 4.8).

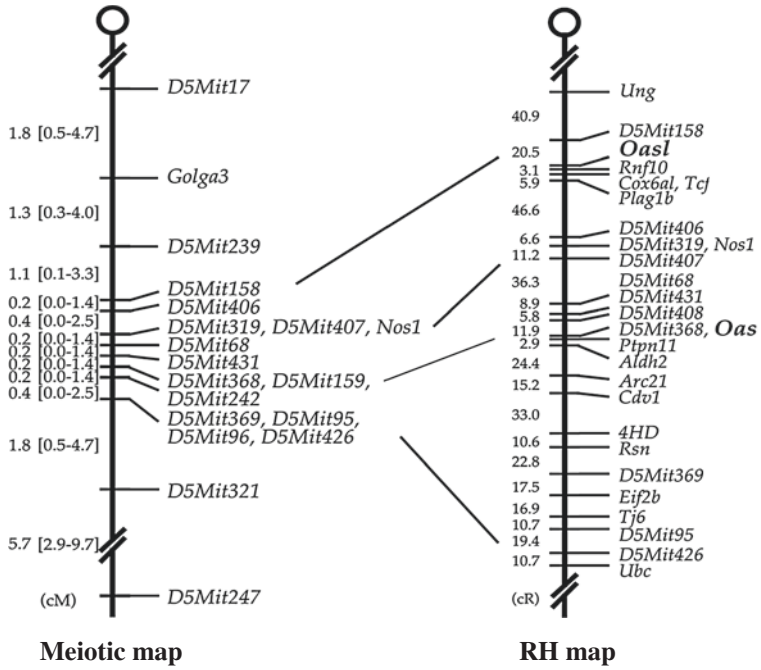


Fig. 4.8 *Merging two independent maps in a consensus map*. These pictures represents two different maps of the regions flanking the locus encoding 2′-5′ oligoadenylate synthetase (mouse Chr 5). The map on the *left* is a meiotic map established by using polymorphic molecular markers (mostly microsatellites) ordered linearly after analysis of the 982 haplotypes of an interspecific backcross progeny (the *EUCIB* resource). The distances are in cM, with 95 % upper and lower confidence limits. Some markers could not be ordered and represent a bin. The map on the *right* was established by using the T31 radiation hybrids panel (*RH* map) and molecular markers. The distances are in centiRays (cR). For most of the markers that are common to the two maps, it can be seen that the order is the same. However, some markers have been ordered in the RH mapping that could not be ordered in the meiotic map (e.g., *D5Mi319-D5Mit407*; *D5Mit369-D5Mit95*). These maps are then complementary and allow the establishment of a refined consensus map of the region (Redrawn from Mashimo et al. 2003)

4.8 Positional Cloning of Mutations and QTLs

The mutations that appear spontaneously in the breeding colonies of inbred strains or those that occur after mutagenic treatment are all interesting, either because they represent potential models of human diseases or, simply, because they can help in the annotation of the mouse genome. For this, however, they must be accurately phenotyped (we discussed this issue in Chap. 2) and at the same time precisely characterized at the molecular level: this is precisely the objective of *positional cloning*.

Positional cloning is a *forward genetic* (from phenotype to genotype) approach whose aim is to characterize the structural alteration at the genome level that is responsible for a specific mutant phenotype.[27] A good historical example of positional cloning is the identification of the gene responsible for the obese mutation (*ob*, now Lep^{ob}) on mouse Chr 6 (Zhang et al. 1994). To achieve this goal, an efficient strategy is to build a high-resolution/high-density molecular map encompassing the mutant locus, then to align and anchor this map on the sequence of the mouse genome that is stored in the databases, with the help of the molecular markers whose sequences are known. This approach is now routine in many laboratories, and it is greatly simplified once a couple of closely linked molecular markers flanking the mutant locus have been identified.

The first step in a positional cloning experiment always consists of setting up a cross in which a great number of polymorphic DNA markers segregates in addition to the mutant allele of interest characterized by a specific phenotype. To set up this cross, it is recommended to select inbred strains that are as distantly related as possible, because this increases the genetic polymorphism segregating in the cross. Similarly, performing an intercross (or F2) between these distantly related strains would be a better choice than setting up a backcross because, in these conditions and as we already commented, two meioses are screened when genotyping each offspring instead of only one in the case of a backcross.

Genotyping a first sample of 50–60 F2 offspring of the cross with the mutant phenotype (equivalent to 100–120 meiosis) for a set of ~80 microsatellites or SNP markers, evenly distributed over the whole genome, is generally sufficient to assign the locus of the mutation into a 20-cM interval (equivalent to 35 Mb), allowing the identification of two markers at the edges of the interval (Fig. 4.9).[28]

Once this first step is achieved, it is then necessary to genotype a much larger progeny (i.e., around 600–800 mice) to yield a greater resolution.[29] Of course, among this large progeny only the mice whose genotype is recombinant in the

[27] *Reverse genetics* is the opposite approach: its aim is to characterize the function of a gene by analyzing the consequences at the phenotypic level of alterations occurring spontaneously or engineered by researchers at the DNA level.

[28] 5–6 markers for the largest chromosomes, 4–5 for the medium-sized and 3 for the smallest is ideal.

[29] The first sample of 50–60 mutant mice generally consists of the first offspring of the larger population.

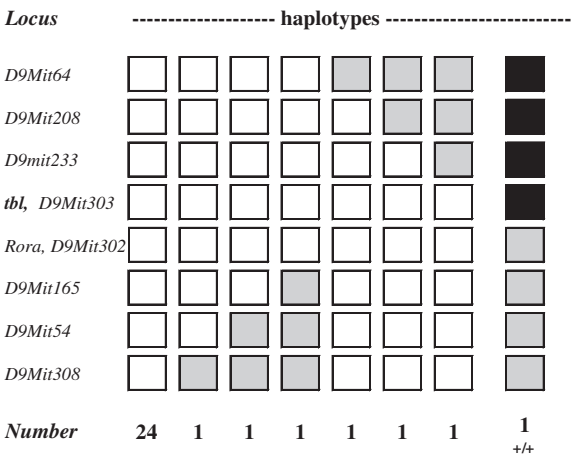


Fig. 4.9 *Positional cloning*. Using an inter-subspecific cross (F2), the mouse mutation *tambaleante* (*Herc1*^{tbl}-Chr 9) was found to map to a 33-cM region of Chr 9 flanked by the DNA markers *D9Mit64* and *D9Mit308*. Analyzing the offspring of a very large inter-subspecific F2, recombinant in the 33-cM interval, allowed separation of the *tbl* locus from the *Rora* locus, a possible candidate gene for *tbl*. The mouse in the *right* column was instrumental for this mapping in the sense that it is recombinant between *Rora* and *tbl*, and proved to be homozygous for the wild-type allele of *tbl* after breeding. This observation was sufficient to reduce the critical interval hosting the mutant allele *tbl* to a 1.6-cM genomic region (Figure redrawn from Mashimo et al. 2009)

interval defined in step 1 have to be genotyped. All others, by definition, are not informative and can be discarded.[30] Once the mapping data are collected, a new molecular map can then be drawn and new molecular markers can be identified that accurately delimit a much smaller critical region where the mutation definitely maps. When the critical interval is in the range of 0.2–0.3 cM (~350–600 kb) or less, no more mapping is necessary and the region can then be inspected in detail for *candidate genes*. Within an interval of 350–600 kb, one expects to find, on average, between 5 and 15 genes whose sequence is available in the different databases.[31]

The last step of positional cloning consists of a functional analysis of candidate genes, for example by asking questions such as: where (in which tissues) are they transcribed? Is the pattern of expression for each of these genes in agreement with the observed phenotype? Are the genes in the interval equally well expressed in normal animals and mutant mice? These basic questions can be answered, for example, by looking at the transcriptional activity of the

[30] This is why it is best to perform genotyping as early as possible, before weaning.

[31] As we will explain in the next chapter, gene density is extremely variable from one genomic region to the next. Accordingly, these estimations must be considered only as indications.

different genes. Another question might be related to the structural integrity of the region: for example, are the PCR amplification products the same size in the mutant and wild-type mice? Many other questions of this kind may be asked, and if we refer to the data collected for the many mutations that have been already cloned by a positional approach, it is likely that two or three genes out of the 5–15 would become top-ranked candidates after this screening. At this step it may then be wise to perform sequencing of the whole region followed by analysis of the data.

Sequencing must be considered as early as possible in a positional cloning experiment because it is a straightforward approach and it is in general sufficient to discover the structural defect causative of the mutant phenotype. Next-generation sequencing (NGS) technology currently offers the possibility of establishing the whole-genome sequence of a mutant genotype for a moderate cost. The defect in question can range from a simple base-pair substitution in an exon or a splicing site to a more extensive rearrangement involving a few kilobases. N-Ethyl-N-Nitrosourea (ENU)-induced mutations are base-pair changes in most cases. Spontaneous mutations are sometimes more complex, with deletions of various sizes.

Difficulties sometimes arise when the sequence comparison points to a missense mutation leading to an amino acid substitution, for example. In this case, which is common after ENU mutagenesis, it is necessary to compare the mutation with the sequence of the wild-type locus or haplotype in the same strain where the mutation arose, because a missense mutation is not automatically correlated with protein dysfunction. In fact, an experiment of positional cloning looks very much like a police investigation, and sometimes it requires time and many redundant clues to unmask the "culprit". There are also cases, fortunately rare, in which the positional cloning does not lead to a conclusion. For example, in the case of the neurological mutation Purkinje cell degeneration (*pcd*), a member of a series of alleles at the *Agtpbp1* locus (Chr 13), its positional cloning ended up inconclusive with no obvious structural change in the coding region or splicing sites (Fernandez-Gonzalez et al. 2002).

4.9 Physical Maps

The genomic DNA of several strains or subspecies of the genus *Mus* has been cloned into a variety of vectors to build *genomic libraries*. Yeast artificial chromosomes (YACs) have been used because they have the advantage of featuring large inserts (500–1,000 kb on average) allowing a reduction in the number of clones in the library. Unfortunately, these vectors have the major drawback of being relatively unstable and unreliable, with chimeric and deleted clones. Bacterial artificial chromosomes (BACs) have been preferred as cloning vectors and have commonly been used in practice. With these vectors, the insert size varies from 80 to 250 kb (Osoegawa et al. 2000).

Using different cloning protocols, different BAC libraries have been prepared with inserts of different sizes, and this has optimized the coverage of the genome by eliminating the gaps due to the cloning protocol. For most of these libraries, the coverage of the genome is high enough to guarantee that any genomic segment has greater than 90 % chance of being represented in at least one clone of the library ($\times 8$ to $\times 12$ coverage).

The extremities of the BAC clones are sequences allowing one to organize the library into groups of head-to-tail overlapping units called *contigs.*[32] Finally, these contigs can be anchored to specific regions of the mouse genome by using the molecular markers (mostly microsatellites) of the high-resolution/high-density map of the whole genome: this results in a *physical map*.

Such a map of the mouse genome was constructed in mid-2002. It was composed of 296 contigs of overlapping BAC clones that were aligned to the human genome sequence on the basis of 51,486 homology matches (Gregory et al. 2002). As we will discuss in the next chapter, this collection of ordered clones encompassing a large part of the mouse genome has provided a framework that has been very helpful for the assembly of the whole-genome sequence. A variety of mouse DNA BACs have been used for making transgenic mice by *in ovo* injection, and it is likely that more mice will be created in the future, with the aim of complementing some of the QTL candidate regions.

4.10 Conclusion

The genetic localization of mouse genes, which started in 1915 with Haldane's discovery of the first linkage group, has been an enthralling enterprise that kept researchers busy for most of the twentieth century and ended with the integral sequencing of the genome. This being accomplished, geneticists have now undertaken the functional annotation of all the genes, including the non-protein-coding sequences. Concurrently, they are associating a biological function to some genomic regions that are not translated into proteins but are nevertheless highly preserved across species. Finally, another important project will be to understand the inheritance of quantitative traits and the structure of the so-called quantitative trait loci (QTLs). All these projects are ambitious and challenging but, here again, the mouse will probably appear to be a privileged model, and many of the tools and strategies that were developed for the purpose of gene mapping (RIS, RCS, molecular markers, etc.) will certainly prove useful.

[32] Because there is *contiguity* in their sequence.

References

Bailey DW (1971) Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. Transplantation 11:325–327

Bateson W, Punnett RC (1906) Comb characters. Rep Evol Comm R Soc Lond II:11–16

Bonhomme F, Selander RK (1978) Estimating total genic diversity in the house mouse. Biochem Genet 16:287–297

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Burgio G, Baylac M, Heyer E, Montagutelli X (2012) Nasal bone shape is under complex epistatic genetic control in mouse interspecific recombinant congenic strains. PLoS One 7:e37721. Epub 2012 May 25

Burgio G, Szatanik M, Guénet JL, Arnau MR, Panthier JJ, Montagutelli X (2007) Interspecific recombinant congenic strains between C57BL/6 and mice of the *Mus spretus* species: a powerful tool to dissect genetic control of complex traits. Genetics 177:2321–2333

Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. Science 250:245–250

Cuénot L (1902) La loi de Mendel et l'hérédité de la pigmentation chez les souris. Arch Zool exp gén 3e séries 3:xxvii–xxx

Darbishire AD (1904) On the result of crossing Japanese waltzing with albino mice. Biometrika 3:1–51

Davisson MT, Akeson EC (1993) Recombination suppression by heterozygous Robertsonian chromosomes in the mouse. Genetics 133:649–667

Davisson MT, Eicher EM, Green MC (1976) Genes on chromosome 3 of the mouse. J Hered 67:155–156

Demant P (2003) Cancer susceptibility in the mouse: genetics, biology and implications for human cancer. Nat Rev Genet 4:721–734

Demant P, Hart AA (1986) Recombinant congenic strains–a new tool for analyzing genetic traits determined by more than one gene. Immunogenetics 24:416–422

Dietrich WF, Miller JC, Steen RG, Merchant M, Damron D, Nahf R, Gross A, Joyce DC, Wessel M, Dredge RD, Andre Marquis A, Stein LD, Goodman N, Page DC, Lander E (1994) A genetic map of the mouse with 4,006 simple sequence length polymorphisms. Nat Genet 2:220–245

Eicher EM (1971) The identification of the chromosome bearing linkage group XII in the mouse. Genetics 69:267–271

Eicher EM (1978) Murine ovarian teratomas and parthenotes as cytogenetic tools. Cytogenet Cell Genet 20:232–239

Eicher EM (1981) Foundation for the future: formal genetics of the mouse. In: Mammalian genetics and cancer: the Jackson laboratory fiftieth anniversary symposium. Alan R. Liss, Inc New York, p 7–49

Eicher EM, Washburn LL (1978) Assignment of genes to regions of mouse chromosomes. Proc Natl Acad Sci USA 75:946–950

Fernandez-Gonzalez A, La Spada AR, Treadaway J, Higdon JC, Harris BS, Sidman RL, Morgan JI, Zuo J (2002) Purkinje cell degeneration (pcd) phenotypes caused by mutations in the axotomy-induced gene, Nna1. Science 295:1904–1906

Flaherty L, Herron B (1998) The new kid on the block—a whole genome mouse radiation hybrid panel. Mamm Genome 9:417–418

Gates WH (1927) Linkage of *Short Ear* and *Density* in the House Mouse. Proc Natl Acad Sci USA 13:575–578

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, Hutton RD, Mullenger IR, Phillips KJ, Smith J, Stalker J, Threadgold GJ, Birney E, Wylie K, Chinwalla A, Wallis J, Hillier L, Carter J, Gaige T, Jaeger S, Kremitzki

C, Layman D, Maas J, McGrane R, Mead K, Walker R, Jones S, Smith M, Asano J, Bosdet I, Chan S, Chittaranjan S, Chiu R, Fjell C, Fuhrmann D, Girn N, Gray C, Guin R, Hsiao L, Krzywinski M, Kutsche R, Lee SS, Mathewson C, McLeavy C, Messervier S, Ness S, Pandoh P, Prabhu AL, Saeedi P, Smailus D, Spence L, Stott J, Taylor S, Terpstra W, Tsai M, Vardy J, Wye N, Yang G, Shatsman S, Ayodeji B, Geer K, Tsegaye G, Shvartsbeyn A, Gebregeorgis E, Krol M, Russell D, Overton L, Malek JA, Holmes M, Heaney M, Shetty J, Feldblyum T, Nierman WC, Catanese JJ, Hubbard T, Waterston RH, Rogers J, de Jong PJ, Fraser CM, Marra M, McPherson JD, Bentley DR (2002) A physical map of the mouse genome. Nature 418:743–750

Grüneberg H (1935) A three-factor linkage experiment in the mouse. J Genet XXXI:157–162

Haldane JBS, Sprunt AD, Haldane NM (1915) Reduplication in mice. J Genet 5:133–135

Hearne CM, McAleer MA, Love JM, Aitman TJ, Cornall RJ, Ghosh S, Knight AM, Prins JB, Todd JA (1991) Additional microsatellite markers for mouse genome mapping. Mamm Genome 1:273–282

Hudson TJ, Church DM, Greenaway S, Nguyen H, Cook A, Steen RG, Van Etten WJ, Castle AB, Strivens MA, Trickett P, Heuston C, Davison C, Southwell A, Hardisty R, Varela-Carver A, Haynes AR, Rodriguez-Tome P, Doi H, Ko MS, Pontius J, Schriml L, Wagner L, Maglott D, Brown SD, Lander ES, Schuler G, Denny P (2001) A radiation hybrid map of mouse genes. Nat Genet 29:201–205

Koseki H, Zachgo J, Mizutani Y, Simon-Chazottes D, Guénet JL, Balling R, Gossler A (1993) Fine genetic mapping of the proximal part of mouse chromosome 2 excludes Pax-8 as a candidate gene for Danforth's short tail (Sd). Mamm Genome 4:324–327

Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet Anal 14(5–6):143–149

Lord EM, Gates WH (1929) Shaker, a new mutation of the house mouse (*Mus musculus*). Am Nat 63:435–442

Love JM, Knight AM, McAleer MA, Todd JA (1990) Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. Nucleic Acids Res 18:4123–4130

Lyon MF (1969) Mapping data. Mouse News Lett 40:26

Mashimo T, Glaser P, Lucas M, Simon-Chazottes D, Ceccaldi PE, Montagutelli X, Desprès P, Guénet JL (2003) Structural and functional genomics and evolutionary relationships in the cluster of genes encoding murine 2′,5′-oligoadenylate synthetases. Genomics 82:537–552

Mashimo T, Hadjebi O, Amair-Pinedo F, Tsurumi T, Langa F, Serikawa T, Sotelo C, Guénet JL, Rosa JL (2009) Progressive Purkinje cell degeneration in tambaleante mutant mice is a consequence of a missense mutation in HERC1 E3 ubiquitin ligase. PLoS Genet 5(12):e1000784

McCarthy LC, Terrett J, Davis ME, Knights CJ, Smith AL et al (1997) A first-generation whole genome-radiation hybrid map spanning the mouse genome. Genome Res 7:1153–1161

Minty AJ, Alonso S, Guénet JL, Buckingham ME (1983) Number and organization of actin-related sequences in the mouse genome. J Mol Biol 167:77–101

Montagutelli X, Serikawa T, Guénet JL (1991) PCR-analyzed microsatellites: data concerning laboratory and wild-derived mouse inbred strains. Mamm Genome 1:255–259

Nijman IJ, Kuipers S, Verheul M, Guryev V, Cuppen E (2008) A genome-wide SNP panel for mapping and association studies in the rat. BMC Genom 9:95

Osoegawa K, Tateno M, Woon PY, Frengen E, Mammoser AG, Catanese JJ, Hayashizaki Y, de Jong PJ (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. Genome Res 1:116–128

Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet 11:221–233

Petkov PM, Cassell MA, Sargent EE, Donnelly CJ, Robinson P, Crew V, Asquith S, Haar RV, Wiles MV (2004a) Development of a SNP genotyping panel for genetic monitoring of the laboratory mouse. Genomics 83(5):902–911

Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, Sargent EE, Asquith S, Crew V, Johnson KA, Robinson P, Scott VE, Wiles MV (2004b) An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 14(9):1806–1811

Petkov PM, Broman KW, Szatkiewicz JP, Paigen K (2007) Crossover interference underlies sex differences in recombination rates. Trends Genet 23:539–542

Pontecorvo G (1976) Polyethylene glycol (PEG) in the production of mammalian somatic cell hybrids. Cytogenet Cell Genet 16:399–400

Rhodes M, Straw R, Fernando S, Evans A, Lacey T, Dearlove A, Greystrong J, Walker J, Watson P, Weston P, Kelly M, Taylor D, Gibson K, Mundy C, Bourgade F, Poirier C, Simon D, Brunialti AL, Montagutelli X, Guénet JL, Haynes A, Brown SD (1998) A high-resolution microsatellite map of the mouse genome. Genome Res 8:531–542

Serikawa T, Montagutelli X, Simon-Chazottes D, Guénet JL (1992) Polymorphisms revealed by PCR with single, short-sized, arbitrary primers are reliable markers for mouse and rat gene mapping. Mamm Genome 3:65–72

Silver J (1985) Confidence limits for estimates of gene linkage based on analysis of recombinant inbred strains. J Hered 76:436–440

Silver J, Buckler CE (1986) Statistical considerations for linkage analysis using recombinant inbred strains and backcrosses. Proc Natl Acad Sci USA 83:1423–1427

Silver LM (1995) Mouse genetics—concepts and applications. Oxford University Press, Oxford

Taylor BA (1978) Recombinant inbred strains: use in gene mapping. In: Morse HC III (ed) Origins of inbred mice. Academic Press, NY, pp 423–438

Williams RW, Gu J, Qi S, Lu L (2001) The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. Genome Biol 2(11):RESEARCH0046. Epub 2001 Oct 22

Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, Graber JH, de Villena FP, Churchill GA (2009) A customized and versatile high-density genotyping array for the mouse. Nat Methods 6(9):663–666

Yokoyama T, Silversides DW, Waymire KG, Kwon BS, Takeuchi T, Overbeek PA (1990) Conserved cysteine to serine mutation in tyrosinase is responsible for the classical albino mutation in laboratory mice. Nucleic Acids Res 18(24):7293–7298

Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM (1994) Positional cloning of the mouse obese gene and its human homologue. Nature 372(6505):425–432

Zou F, Gelfond JA, Airey DC, Lu L, Manly KF, Williams RW, Threadgill DW (2005) Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. Genetics 170(3):1299–1311

Chapter 5
The Mouse Genome

5.1 Introduction

In Chap. 4 we explained how mouse geneticists were able to develop high-density and high-resolution genetic maps of the mouse genome by taking advantage of the unequaled strategies and tools they had at their disposition: i.e., inter-subspecific crosses, recombinant inbred strains, radiation hybrids and a wealth of polymorphic molecular markers of all kinds. We also explained how the same geneticists could develop physical maps by anchoring virtual (i.e., in silico) DNA fragments cloned into BACs, YACs or cosmids onto the molecular markers previously ordered along each chromosome. It is clear that, while building these maps and associated libraries of cloned DNAs, geneticists were in fact gathering the essential ingredients for undertaking the logical next step: the sequencing of the whole mouse genome.

The decision to undertake such an ambitious (and, at the time, expensive) project was made at the turn of the millennium and was strongly influenced by the decision to sequence the human genome, made a few years earlier (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). A first draft of the mouse genome sequence was released in 2002, only a few months after the release of the first draft sequences of the human genome (Mural et al. 2002; Mouse Genome Sequencing Consortium, Waterston et al. 2002) and 2 years before the publication of the rat sequence (Gibbs et al. 2004).

The completion of these projects, as we will see in this chapter and the following chapters, had an enormous impact in many areas of genetics and biology. Making these genome sequences available to the community provided a wealth of information about genome structure and evolution through the identification of similarities and differences across species. As Robert Waterston and his colleagues wrote in the conclusion of their publication: "*The mouse provides a unique lens through which we can view ourselves* […]. *With the availability of* [its genome] *sequence, it* […] *provides a model and informs the study of our genome as well*" (Mouse Genome Sequencing Consortium, Waterston et al. 2002).

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_5

Nowadays, geneticists have direct and free access to a variety of high-quality genomic sequences through the Internet, and most of them would probably find it difficult to work without having these tools at hand.

5.2 The Sequence of the Mouse Genome

The availability of the mouse genome sequence represented such an important piece of information for the development of the genetics of this species that it would certainly have become available sooner or later, for example, as a consequence of the continuous addition of the ever-increasing number of sequence fragments released by independent laboratories. However, such a disorganized approach would have inevitably resulted in delay, in a sequence with plenty of gaps and redundancies, and finally in a higher cost. Retrospectively, the decision to give support and priority to the complete and systematic sequencing of the mouse genome, and to make it a concerted project completed by a team of specialists, should be considered very wise. This decision was also very altruistic because the laboratories that did not have easy access to sequencing facilities, for whatever reasons, can now benefit from this resource, entirely free of charge, for designing their experiments. Further evidence of this achievement is provided by the enormous and ever-increasing number of scientific papers that have been published since the release of the initial draft sequences of the mouse genome and make direct reference to it. This trend will certainly grow in the years to come with the progress made in sequencing technologies and the associated dramatic reduction in cost.

The sequence of the rat genome has also turned out to be a valuable piece of information for geneticists, because it has allowed three-way comparisons with the other two species (human and mouse). These comparisons have provided details about how evolution proceeds over a relatively short timescale. As mentioned in Chap. 1, the human and rodent lineages split around 75–80 Myr ago, while the mouse and rat lineages split around 12–14 Myr ago.

5.2.1 The Mouse Genome is Enormous in Size, and its Structure is Complex

Measurements of the intensity of the brilliant purple color performed on mouse cell nuclei (early spermatids, for example), after a Feulgen reaction, indicated that the DNA content of the mouse haploid genome corresponds to approximately 3×10^{-12} g (= 3 pg), which translates into a molecular weight of ~1.8×10^{12} daltons (Da). Since the average molecular weight of a double-stranded DNA base-pair (bp) is ~600 Da, this means that one expects to find ~3×10^9 bp or 3.0

Giga-base-pairs (Gb) of DNA in a mouse haploid genome (Silver 1995). This is ~650 times more than in the genome of the bacterium *Escherichia coli* K-12, which comprises 4,639,221 bp. To translate this into more palpable terms, we computed that, if the haploid mouse DNA sequence was printed as a single line using the 11-point `Courier` font, all in uppercase, to symbolize the four bases (`A`, `T`, `G`, `C`), the length of this line would be roughly equal the distance from London to New York City (5,600 km or 3,480 miles). To express this still differently, the printed transcription of the message in 12-point Times font would represent around 3,500 books with a size similar to the one you have in your hands. However, although obviously enormous, this sequence can be stored on the hard disk of a personal computer (Silver 1995). Finally, mouse nuclear DNA has an A + G/C + T ratio of 49.99/50.01 (~1), as in human.

Aside from its large size, the mouse genome is also heterogeneous. Biophysicists who studied the thermodynamics of nucleic acid denaturation/renaturation had already recognized this peculiarity, over 40 years ago, by measuring the $C_0t_{1/2}$ value, a parameter reflecting the structural heterogeneity of a DNA sample that is based on the speed of reconstitution of double-stranded DNA (dsDNA) from previously denatured single-stranded DNA (ssDNA). The same biophysicists also demonstrated that some fractions of the mouse genomic DNA renatured much faster than others as a consequence of a high proportion of repeated sequences.

Another interesting comparison is between the physical size of the mouse and pufferfish (*Takifugu rubripes*) genomes, leading to the observation that the genome of the fish is about nine times smaller (0.35 pg of DNA or 340 Mb) than that of the mouse. Considering that all vertebrates presumably have a similar number of protein-coding genes (between 20,000 and 30,000, as we will discuss further), it has been suggested that the difference in size between the two genomes is probably due to the presence, in the mouse but not in the fish, of non-protein-coding DNA sequences of unknown function.

The mouse genome also contains sequences that are repeated many times. This was revealed by the observation that, if we use a randomly cloned 1–2-kb DNA segment as a probe and label it with a fluorescent dye, in most cases this probe will hybridize with several chromosomal regions, indicating extensive redundancies.

Finally, if we consider that there are between 20,000 and 30,000 genes in a mouse genome (which is a reasonable guess) and only 4,377 genes in *E. coli*, this indicates that the average gene density in the mouse is much lower than in the bacterium (~1/100 kb in the mouse versus roughly ~1/1 kb in the bacterium). All these observations support the idea that a large proportion of the mouse genome does not code for proteins and may represent what Susumu Ohno called "*junk*" DNA (Ohno 1972)—unless we find that part of the DNA in question serves other functions that might be important.

Considering all these issues (i.e., a genome with a large size, with a heterogeneous structure, with many redundancies and a large amount of possibly "*junk*" DNA), scientists were then warned from the beginning that unraveling the complete sequence of a mammalian genome would be a long and difficult enterprise.

5.2.2 How Was the Mouse Genome Sequenced?

There are basically two strategies for sequencing a complete mammalian genome. The first one, known as *hierarchical shotgun sequencing* (HSS), makes use of cloned DNA with large inserts such as bacterial artificial chromosomes (BACs—with 150–250 kb DNA inserts), P1 phages or, less frequently, yeast artificial chromosomes (YACs—200–1,000 kb). As explained in Chap. 4, these clones of DNA are assembled into a series of overlapping elements known as *contigs* (from contiguous DNA segments), which altogether make a physical map encompassing chromosomal segments of the greatest possible dimension. The DNA clones mentioned above are generally selected once they have been thoroughly checked for structural integrity, rejecting those that are chimeric or have deletions (a situation that is common in YACs but less common with BACs). The assembly of these cloned DNAs into contigs is achieved by careful fingerprinting of each and every clone. When the contigs are established, in general from several individual clones ranging from 100 to 1,000 kb, a subset of minimally overlapping clones is chosen and each of its elements is sequenced several times to minimize the effect of sequencing errors (this minimal set is sometimes called the "*Golden Tiling Path*" or simply the "*Golden Path*"). The primary sequence is called a *read* and the released genome sequence, or *draft*, results from the integration of several independent reads (in general 10–15, sometimes more). After computerized processing of these independent reads, and if we suppose that the sequencing errors occur randomly, the final rate of errors in a given consensus sequence is very low, in general less than one error per 10^5 bp.

The HSS strategy is relatively slow and tedious, but it is systematic, progressive and highly reliable. The use of clones with large DNA inserts is also a way to circumvent, at least to a certain extent, the difficulties associated with the sequencing of DNA repeats and variations in copy number, which are true nightmares for sequencers. However, the HSS strategy has the disadvantage that only long DNA fragments cloned in a vector can be sequenced. Unfortunately, it is virtually impossible to clone the whole of a mammalian genome in BAC or YAC vectors, for reasons that are associated with both the structure of the DNA in some chromosomal regions and with the cloning technology.

A second strategy, called *whole-genome shotgun* (WGS), consists of the mechanical fragmentation (e.g., by sonication) of the mammalian DNA into segments measuring 100–400 bp, which are sequenced from both ends using the *chain termination method*. Multiple reads of the targeted DNA are obtained by performing several independent rounds of this fragmentation, each followed by sequencing. Once the sequence of the targeted DNA is achieved, computer programs are then used to assemble the pieces of the puzzle, ordering the individual fragments into virtual contigs, then in super- or hypercontigs and finally in ultracontigs based on the overlapping sequences of the different reads.

The WGS method is fast and (in theory) does not require the pre-existence of a physical map. Unfortunately, it does not allow the sequencing of certain genomic

segments such as highly repeated regions. Combining the two strategies (WGS first, then HSS) allows for the correction of most of these difficulties. In short, the two strategies are complementary: WGS provides rapid and relatively good coverage early in a project, while HSS is more systematic and more efficient for the sequencing of regions with repeated sequences. The human genome was sequenced by using mostly the HSS strategy, while the mouse and all other mammalian genomes were sequenced by using mostly the WGS strategy, with the help of HSS only for finishing some regions.

In fact, technical and methodological difficulties emerge when the objective is to sequence the genome of a species for the first time (the human genome in this case), but the situation is greatly simplified when the project is to sequence the genome of evolutionary related species. This is because it is possible to take advantage of the existence of the many interspecific structural homologies that exist at the chromosomal level. Thus, the mouse and rat genomes were sequenced mostly by WGS, and accordingly were completed much faster than the sequencing of the human genome (Fig. 5.1).

Sequencing techniques have progressed enormously recent years and many steps are now fully computerized, reducing human intervention and cost. The latest assembly released by the Mouse Genome Sequencing Consortium (MGSC) has a length of 2,730,871,774 bp (Golden Path from *Ensembl*—September 2013). Curators of the database consider that at least 99 % of the mouse genome sequence is established, with the exception of only a few small gaps (~180) scattered in between a total of 750 contigs, with less than one sequencing error per 10^5 bp. All of the chromosomes have been entirely sequenced, including the X and the Y, allowing comparisons with homologous regions of the human and other mammalian genomes to be performed at a very high resolution.[1]

Such comparisons, revealing similarities and differences, are a rich source of information. Similarities (i.e., sequence conservation), as we will discuss later, allow us to detect regions that are very likely under selective pressure and which, for this reason, have remained unchanged or nearly so for millions of years, indicating that they are presumably genetically important and, accordingly, have resisted random drift. Differences at the sequence level may be even more interesting a priori, because they may contain information explaining how speciation proceeds. It will be obviously interesting to discover both the mechanisms governing these processes and the consequences of these differences at the phenotypic level. We will come back to this point several times, which is well exemplified in the case of variations in gene or DNA copy numbers (copy number variations or CNVs, see Sect. 5.3.6.).

The mouse sequencing project was undertaken by the MGSC, an organization that consisted originally of three laboratories: the Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology (USA), the Washington University Genome Sequencing Center (USA), and the Wellcome Trust Sanger

[1] The mitochondrial DNA has also been sequenced. See Sect. 5.6.

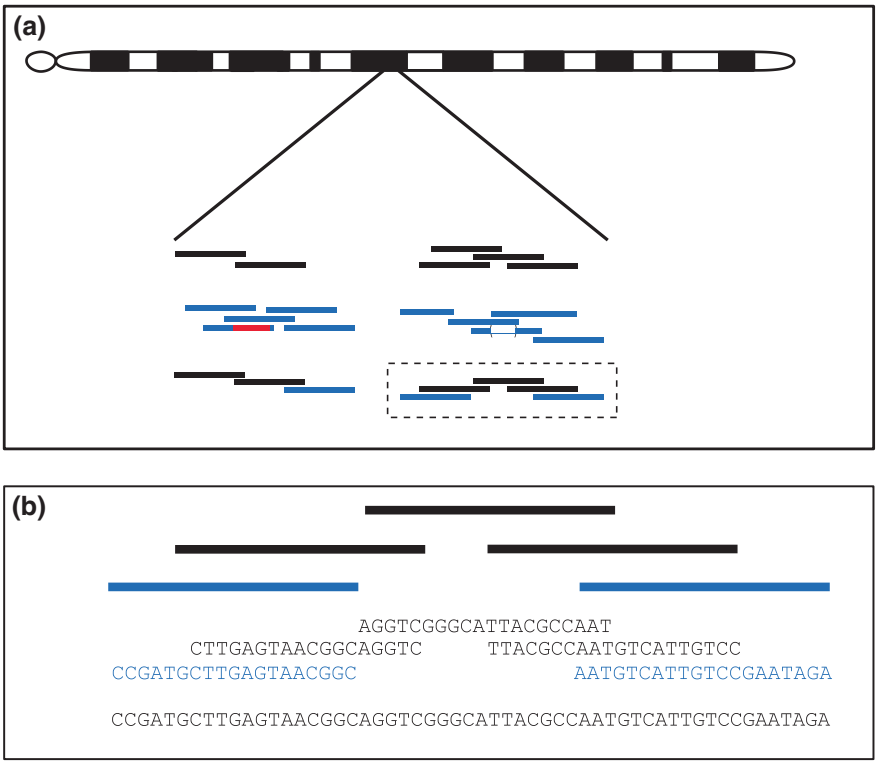


Fig. 5.1 *Strategies used for sequencing mammalian genomes*. Two strategies have been used for sequencing the mammalian genomes: hierarchical shotgun sequencing (HSS) and whole-genome sequencing (WGS). HSS (Fig. 5.1a, b) has been used for sequencing the human genome. It works in two successive steps and makes use of bacterial artificial chromosomes (BACs, ~150–300 kb) or yeast artificial chromosomes (YACs, ~500–2,000 kb) that have been previously used for the establishment of the physical map or "contig map". In the first step (**a**), the integrity and quality of these cloned DNAs is carefully checked (absence of mosaicism, absence of deletion). Then the most interesting elements (**b**) of these contigs (those representing the "*golden path,*" with minimum overlap) are completely sequenced and the sequence ordered. The HSS strategy is systematic and reliable, but it is slow and does not allow the sequencing of regions with repetitive DNA. The whole-genome sequencing strategy (WGS) (Fig. 5.1c, d, e) has been used for sequencing most of the mouse genome. This strategy completely bypasses the BAC/YAC step and consists of the direct mechanical fragmentation of DNA samples to obtain a mixture of independent, randomly cut stretches of DNA 100–400 bp long (**c**). These stretches are then cloned using adaptors, labeled, and sequenced end-to-end (**d**). In a third step (**e**), sequence overlaps are looked for by using appropriate computer software and the clones are then arranged in a head-to-tail manner to form virtual contigs of non-redundant, top-quality sequences. In the final step, the contigs are anchored to the specific chromosome they belong to. The process is generally repeated several times to reduce the number and size of the unsequenced regions and strengthen the quality of the sequence. The gaps in the sequence resulting from the WGS strategy are filled, where possible, by HSS. In the current mouse sequence, the number of gaps is extremely reduced

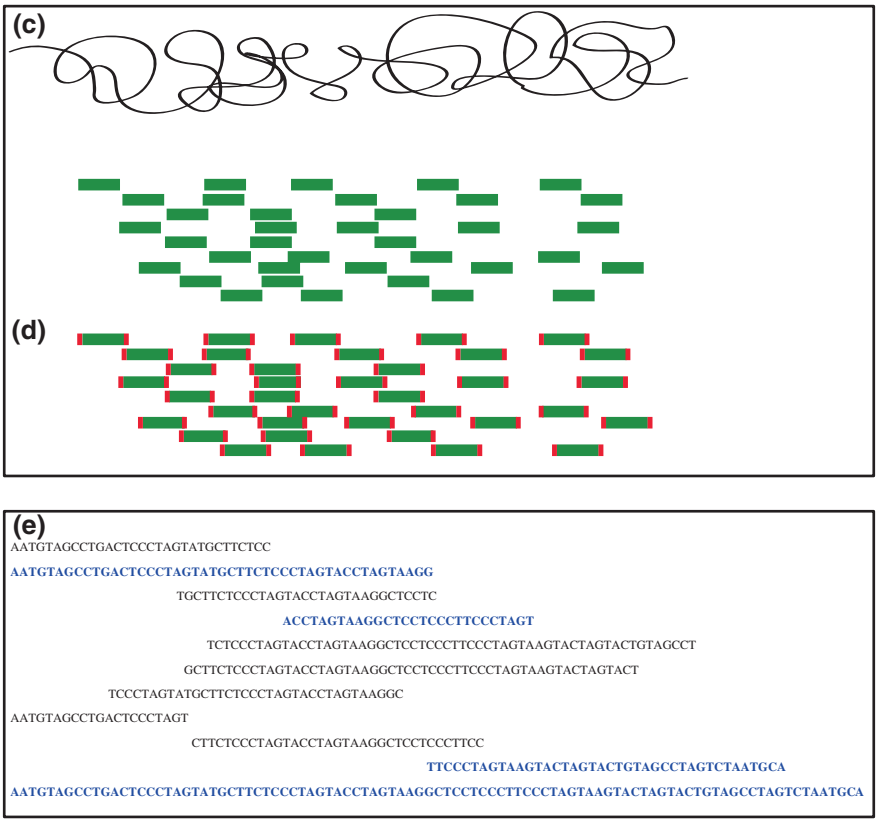


Fig. 5.1 (continued)

Institute (UK). Based on discussions with the scientific community, MGSC investigators decided to sequence, first, the genome of a female from the C57BL/6 inbred strain. At the same time, four other inbred strains (A/J, DBA/2J, 129X1/SvJ, and 129S1/SvImJ) were being sequenced by the CELERA firm in another independent WGS project. Here again, interstrain comparisons have been of great interest when matched with particular phenotypes. Nowadays, the original projects are finished, even though molecular biologists at the MGSC keep working on some specific regions. The Mouse Genomes project from The Wellcome Trust Sanger Institute recently completed the sequencing of an additional 17 inbred mouse strains: 129P2, 129S1/SvImJ, 129S5, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CAST/EiJ, CBA/J, DBA/2J, LP/J, NOD/ShiLtJ, NZO/HiLtJ, PWK/PhJ, SPRETUS/EiJ, and WSB/EiJ (see http://www.sanger.ac.uk/resources/mouse/genomes/). These strains were very carefully selected after extensive discussions via the Internet among the members of the community of mouse geneticists. The genome of the FVB/N inbred strain, popular for the production of transgenic animals and for skin carcinogenesis studies, is now also available (Wong et al. 2012).

These genome sequencing projects are now benefiting from new, ultra-efficient sequencing technologies known as *next-generation sequencing* (NGS). It is likely, for example, that many genome sequences from highly informative strains (strains from the Collaborative Cross, for example—see Chaps. 9 and 10) or even some carefully selected individual mice will become available, contributing efficiently to the analysis of complex traits. Even if the development of bioinformatics resources for the interpretation of the tremendous and ever-increasing amount of data remains a challenge, we can say that the mouse genome-sequencing project is, without any qualification, a complete success from an analytical point of view. However, from now on scientists will have to consider a new challenge, at least as important: the annotation of all sequences in this genome. No doubt they will be kept very busy for another few years.

5.3 The Structure of the Mouse Genome

Once a genome is entirely sequenced and the sequence stored in a database, scientists can then start looking at it in more depth. This structural analysis, run in parallel with a functional analysis, is part of the so-called *genome annotation process*, and one of the first challenges in this matter is to identify and characterize as accurately as possible the DNA regions containing the genes proper (i.e., the DNA coding for proteins or RNAs), the regulatory elements, and some other potentially important structures. This is a real challenge because, if we recall what we said earlier when discussing gene density in mammalian genomes, the protein-coding and related sequences represent only a very small proportion of the mammalian DNA. However, if we consider that this functionally important fraction of mouse DNA, because it is under the constraint of *purifying* (i.e., negative) *selection* during evolution, is likely to be highly preserved across different species, we already have outlined a strategy to identify and estimate it. This estimation has been achieved, shortly after the release of the first draft of the mouse sequence, by cross-comparing several regions of the human genome with various short sequences of the mouse genome, and the answer was that there is indeed great interspecific homology (over 95 %) for around 3.5–5 % of the genomic DNA sequences. There are good reasons to believe that the genes encoding proteins and other important sequences are gathered in this fraction.

5.3.1 Finding the Coding and Related Sequences

The first step in the process of genome annotation generally consists of checking for the presence or absence in the newly sequenced genome of some specific sequences previously characterized in other species (the exons, for example), and

evaluating the number of copies, their organization and flanking sequences, etc. The geneticist may also wish to make an inventory of all the genes of a given species: those encoding proteins and those transcribed only into RNAs. These questions have triggered a multitude of intensive studies, many of which have now resulted in more or less precise answers.

5.3.1.1 Retrieving Specific Sequences

Nowadays, finding a particular sequence in a genome is relatively easy and several software packages have been designed for this purpose. The most popular is BLAST. BLAST allows similarity searches to be performed against any databases of recently sequenced organisms. BLAST will rapidly identify and retrieve a sequence in the human or rat genome that resembles a mouse sequence based on similarity of sequence. These software packages work, roughly, like the sub-programs that are activated when, working on a text file, one selects the command "*Find*" to specifically retrieve a chain of characters, with the important difference that BLAST can retrieve sequences that are not 100 % identical to the queried one. ROSETTA[2] and SEQUENCHER® sequence analysis softwares are other packages useful for finding genes (and not only coding sequences) by comparisons, for example, between human and mouse DNAs. ROSETTA performs sequence alignments and compares the exon sizes, splicing sites, etc., and finally makes gene predictions.[3]

When a coding sequence (a mouse exon, for example) is used as a template for retrieving the most closely related sequences in the human or rat genomic sequence, in ~95 % of the cases BLAST retrieves a sequence with high similarity and 90 % of these sequences are on the homologous chromosomal segment in all three species. Geneticists say that they share the same *syntenic* location (from the Greek, meaning "on the same ribbon") and these genes are called 1:1 *orthologs*. This indicates that most of the genes in a given mammalian genome are part of an ancestral heritage and do not vary much among other mammalian species even if, sometimes, there are variations in terms of copy numbers, as we will discuss further. Differences in terms of presence versus absence are rare but occasionally occur. For example, approximately one hundred predicted mouse genes identified in the initial mouse draft sequence were reported as missing (having no homologous counterpart) in the human genome. The reverse of course is also true, and some human genes are absent in both the mouse and rat genomes. A good example of such a situation is the gene encoding human interleukin 8 (*IL8*), which cannot be found in the rat and mouse regions of homology for HSA-Chr 4 (see Fig. 5.2).

---

[2] https://www.rosettacommons.org/.

[3] Sequencher version 5.1 sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA http://www.genecodes.com.

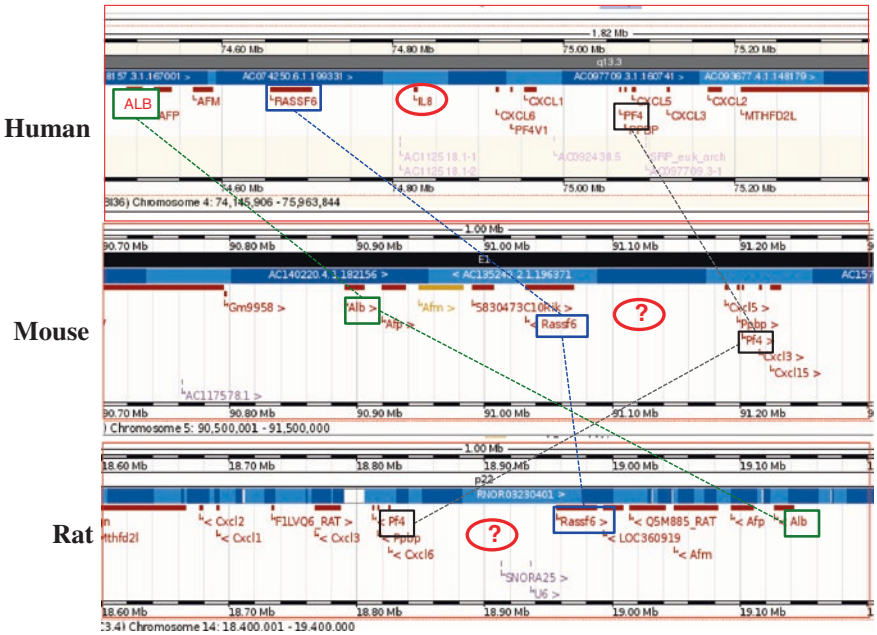


Fig. 5.2 *Sequence comparisons between mammalian genomes*. The orthologous copy of the human gene encoding interleukin-8 (*Il8*) is missing in the mouse and rat genomes. The figure shows the region of human chromosome 4 (HSA4) where the *IL8* gene is located, with the homologous regions in mouse chromosome 5 (MMU5) and rat chromosome 14 (RNO14). The rat chromosome is affected by a paracentric inversion when compared with the human and mouse homologous regions. Such rearrangements are extremely common in the mammalian genomes and are very useful (with other methods) for establishing the phylogenetic relationships between species. The images are from the *Ensembl* Genome Browser database (May 2013)

These qualitative differences are not easy to explain and can result either from true deletions, with no consequences at the phenotypic level, or from the fact that the supposedly deleted genes in fact still exist elsewhere in the genome but have evolved so rapidly, in one or the other lineage, that they are no longer recognizable as orthologs based on sequence comparisons. The first hypothesis is the most likely, since such segmental deletions of recent origin have been discovered, by chance, in the genome of several inbred strains while others were reported normal (undeleted). For example, mice of the C57BL/6JOlaHsd substrain (also known as C57BL/6S) are homozygous for a deletion encompassing the entire α-synuclein gene (*Snca*-Chr 6) (Specht and Schoepfer 2001). These mice are fertile and have a normal lifespan, but they have at least one gene inactivated compared with most other C57BL/6 substrains. Examples of this kind have been reported in many other laboratory inbred strains and also exist in the human and rat species (Perez et al. 2013).

Finding genes in the genome of one species, once the orthologous versions of these genes are known and already identified in the genome of another closely

related species (such as human, rat, and mouse), is then relatively straightforward and many computer programs can do this, even if surprises and difficulties occasionally occur, as we will see later.

5.3.1.2 Identification of the Coding Sequences

The situation is more complicated when the objective is to identify all the coding sequences (all the exons, for example) in a freshly sequenced genome.

A first and relatively efficient technique, known as *exon trapping*, was published in 1991 (Buckler et al. 1991). With this technique, a cloned genomic DNA was inserted, by genetic engineering, into an intron of the human immunodeficiency virus 1 (HIV-1) *tat* gene (Trans-Activator of Transcription), contained within the plasmid pSPL1. COS-7 cells were then transfected with these constructs, and the resulting RNA transcripts were processed in vivo. The splice sites of exons contained within the inserted genomic fragment were put in phase with the splice sites of the flanking *tat* intron. The mature RNA collected from the COS-7 cells contained the potential exons, which could then be amplified via RNA-based PCR and ultimately cloned.

Exon trapping has been a very helpful technique, especially in the projects whose aim was the positional cloning of a gene identified only by a mutant allele. However, compared to more recent techniques, it has two major drawbacks: (*i*) it does not trap faithfully the large or very small exons, and (*ii*) it is relatively expensive because it requires a significant amount of bench work and in vitro cell culture.

5.3.1.3 Using Expressed Sequence Tags (ESTs) for the Detection of Transcribed Sequences

Taking into account the fact that several mammalian genomes are now entirely sequenced, strategies have been developed that are based on the identification at the genome level, by all possible techniques, of sequences deduced from transcribed products. One of the first strategies consists in using so-called *Expressed Sequence Tags* (ESTs). ESTs are short sub-sequences of cDNA corresponding to a few hundred (~350–500) base-pairs of a cDNA, starting from the 3' end, sometimes from the 5' end. Millions of such ESTs (from several mammalian species) are available in public databases, and the sequence of each of these ESTs can be used as a molecular probe to retrieve the complete sequence of the gene the EST belongs to (or is related to), simply by "pulling on" the flanking sequences. Since the ESTs stored in a given database were in general prepared from a specific tissue (brain, blood, skin, neoplastic tissue, etc.) at a certain step of development (embryonic, 10 days, adult, senescent, etc.), using these ESTs for gene identification has the additional advantage of providing information concerning the transcriptional level and the gene expression pattern for the annotation process. ESTs have been

instrumental for the initial identification of many genes in the mouse as well as in the human genome, and still are. In addition, the sequence alignments can be performed entirely in silico, which means rapidly and at virtually no cost. The major drawback of these ESTs is that only a fraction of the genes are expressed simultaneously, and consequently the EST collection in a particular database represents only a fraction of the genes of a given species. Finally, some genes are transcribed only in particular circumstances, at very low levels, or transiently and, by definition, they are poorly represented in EST libraries or databases.

5.3.1.4 Using Strategies Based on Artificial Intelligence

Other strategies, requiring sophisticated informatics, rely on the identification of some transcription-related motifs that are part of most protein-coding genes (Blanco and Guigo 2005; Harrow et al. 2009) (see also next section). These motifs have been successfully used for gene detection with software systems like GENSCAN, developed by Burge and Karlin (1997). In addition to the strategies mentioned above, more refined prediction programs, often referred to as de novo or ab initio *gene finders*, have also been developed by geneticists and computer scientists. These programs are based on the existence of subtle differences at the sequence level that can be used to sort out putative coding regions from non-coding regions by making use of the so-called hidden Markov chain models. These prediction models are based on the fact that biases and dependences exist in coding sequences that are not observed in non-coding regions. This means, for example, that the five preceding bases influence the probability of finding a particular base at the sixth position of a new sequence if, and only if, the sequence in question is a coding sequence. When scanning a novel nucleotide sequence, the program computes a *coding likelihood score*, based on a Markov chain model of order 5, and makes an assessment as to whether the sequence is more likely to be from an intron, exon or intergenic region (Harrow et al. 2009).

All these sequence prediction algorithms are being constantly improved based on the experience acquired from training with DNA samples whose sequence is fully annotated. These programs work more or less like the software designed for language translation. Years ago, the meaning of "computer-translated" sentences was only remotely related to the meaning in the original sentence and sometimes limited to an unordered set of key words. Nowadays, the quality of the translation is very good (at least for certain languages). Based on their encouraging results, researchers consider that, as of today, around 85 % of genes should be rapidly and easily detected in any new mammalian genomic sequence by using software of this kind. Most of these newly discovered genes must, however, be validated by other approaches because the discovery of a gene-like structure does not automatically mean that an authentic, indisputable, and functional protein- or RNA-encoding gene has been "fished". This validation is very important work, whose aim is to create a gold-standard reference for gene annotation. A program of this

kind has been undertaken by the Human and Vertebrate Analysis and Annotation (HAVANA) team at the Sanger Institute, where the human, mouse, and zebrafish genomes are carefully annotated manually.

Making sequence comparisons (or alignments) with other genomes (human, rat, zebrafish) has allowed a rather rapid identification of a great number of mouse genes. However, from now on, the identification of novel genes in the mouse will probably progress at a somewhat slower pace because the situations researchers face are sometimes difficult. Some genes, for example, are very large and extensively fragmented, while others are very small with only one intron or even no intron at all (for example, the intronless genes encoding RNAs and histones). Since neither of these two categories of genes correspond to the "canonical" representation of most mammalian genes, they have to be annotated manually and this takes much more time. Another very common situation is that, although they share a syntenic location as expected, orthologous genes are not always in a 1:1 ratio but rather in 1:2, 1:3, and so on. We will describe situations of this kind, where the "pseudo-orthologous" copies are sometimes slightly altered or incomplete, but are still transcribed and accordingly annotated as a true gene.

Finally, overlapping and nested genes have been shown to exist in mammals just like in *Drosophila,* with various imbrications of their structure with their neighboring genes. Nested genes were generally described as genes with a relatively short size, consisting in general of only one exon and entirely nested within a single intron of a host gene. The situation has recently changed dramatically as a consequence of more in-depth analysis of the mouse transcriptome, as we will discuss further in this chapter, and many RNAs are transcribed from the mouse genome whose function is not yet established. In the same way, genes have been found that are transcribed in the opposite orientation to their neighboring host genes, and sometimes negatively influence the transcription of these genes via antisense-mediated inhibition (see below—Chap. 6 on X-inactivation). Identification of nested genes is difficult but, fortunately, approximately 60 % of nested genes are conserved in mouse and human in the same genomic context.

The ENCODE project (ENCyclopedia Of DNA Elements), which is essentially the next step for the Human Genome Project, has set as its major aim the establishment of all the structural and functional elements of the genome. It is definitely an ambitious project but it makes a lot of sense and is really necessary if we consider its potential applications. Here again, just like for the sequencing of the mouse genome, we can say that this meticulous analysis conducted at the DNA level would have to be achieved one day because the general feeling of the community is that it is a crucial endeavor, if not simply the essence of genetics: then why not do it right now, as rapidly as possible, on a systematic basis?

The preliminary results of the ENCODE project, although still fragmentary, have already changed our understanding of the mammalian genome by demonstrating that the mammalian DNA hitherto labeled "*junk*" might not be *junk* after all.

5.3.2 The Canonical Architecture of a Protein-Coding Gene

As discussed in the preceding section, many points remain to be elucidated concerning the structural organization of the mouse genome. However, as of today, hundreds of genes have been entirely sequenced in several species including mouse, rat, human, and domestic animals. As a result, it is now possible to outline the classical or canonical architecture of the "average" mammalian gene.

A gene is a segment of DNA that encodes an RNA molecule that may or may not be translated into a protein. For this reason, geneticists formally distinguish two types of genes: the protein-coding genes and the non-protein-coding genes. For many years, and up to relatively recently, molecular geneticists considered that the two strands of the DNA molecule were not equivalent: one of them was the coding strand while the other was the template or anticoding strand. However, and unexpectedly, it has recently been demonstrated that mammalian DNA is pervasively transcribed from both strands. We will come back to this important point later in this chapter when discussing the transcriptome and the non-protein-coding RNAs. Here, we will simply discuss the organization of a classical protein-coding gene as it has been established as the result of thirty years of careful positional cloning, sequencing, and annotation.

The transcription of a protein-coding gene into a primary mRNA proceeds from the 5' to the 3' end and starts ~50–60 bp upstream of the first AUG codon, encoding a Methionine. The ~50 bp between the transcription initiation site and the initial AUG is part of the so-called 5'-untranslated region (5'-UTR) or *leader sequence.* This sequence usually contains a ribosome binding site (RBS), known as the *Kozak sequence* (gcc)gcc(A/G)cc**AUG**G (Kozak 1987), that includes the AUG initiation codon.

Upstream of the transcription start site at the 5' end, several consensus sequences have been identified that are part of the promoter sequence of the gene, as we will see in the next section of this chapter. Opposite to the 5'-UTR is the 3'-UTR or *trailer sequence,* required for the processing of mRNA, the size and canonical sequence of which is not as precisely known as that of the leader sequence. The end of a structural gene is called the *transcription termination site*. Some specific sequences are also found in the 3'-UTR. First is a *polyadenylation signal* composed of sequences like AAUAAA or a slight variant. The polyadenylation signal indicates that transcription will be terminated approximately 30 base-pairs downstream of it, while a tail composed of a few hundred adenine residues (the poly-A tail) will be added to the transcript. The poly-A tail is important for the nuclear export, translation, and stability of the mRNA.

Since 1977, it has been established that many (around 60 %) mammalian genes have a heterogeneous structure: some parts are included in the final protein or RNA product, while others have another destination or are merely degraded. Hence, the coding sequences of most mammalian genes are composed of an alternation of exons (*ex*pressed regi*on*) and introns (*intr*agenic regi*on*). Introns are spliced off during RNA processing or maturation when the pre-mRNA becomes

a mature mRNA, ready to be translated into a protein product. RNA splicing is a complex and very precise procedure that is regulated and controlled at the cellular level, at the base-pair level of precision (Fig. 5.3). This process requires several highly specific tools: at least five small nuclear RNAs and around 150 proteins, collectively known as the *spliceosome* (Hoskins and Moore 2012). Among the most important are the small nuclear ribonucleic acids or snRNAs, the small nucleolar RNA or snoRNAs, and specific enzymes including the ribonucleoproteins or "snurps".

Splicing sites can be identified at the DNA level because they have a consensus sequence: the first two bases at the beginning of an intron (at the 5' end) are almost always GT and the last two, at the 3' terminus of the same intron, are almost always AG. The sequences immediately upstream of the AG and downstream of the GT are also conserved, although to a lesser degree. For example, the intronic region upstream from the AG is usually a region rich in C and T. The regions at the 5' boundaries of the introns are called the *donor sites* and those at the 3' end are called the *acceptor sites*. We already mentioned earlier that these splicing sites have been used for the identification of exons with the exon trapping technique. Most sequence identification software can also identify these sites in a mouse DNA sequence and label them as "candidate splicing sites" (Fig. 5.4).

Not all exons in a gene are spliced and subsequently assembled to form the final RNA product. In fact, if we consider that the exons correspond to "functional units" and the introns are "spacers" between these functional units, we observe that the exons can be assembled into different combinations to produce different polypeptides. This is known as *alternative splicing* and it is estimated that ~95 % of multi-exonic genes are alternatively spliced in mammalian genomes. From numerous observations, it is also known that the exons in a gene are of two types: (i) those that are always present in all transcripts, which are often referred to as *constitutive or major forms of transcripts;* and (ii) those that are *optional* or *alternative*. Exons of the second type, those that are only included in some spliced

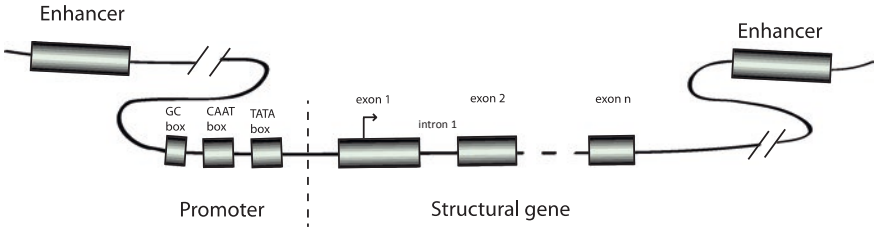


Fig. 5.3 *Canonical (and simplified) representation of a protein-coding mammalian gene*. The enhancers represented in the figure are not always present and are sometimes distant from the promoter region by several Mb. Many sequences in the promoter region are important for gene regulation, but not all of them have been identified and they probably vary from one gene to another. Not all genes have a CAAT box or a TATA box. Finally, not all genes have intronic sequences, and not all exons are represented in the final product

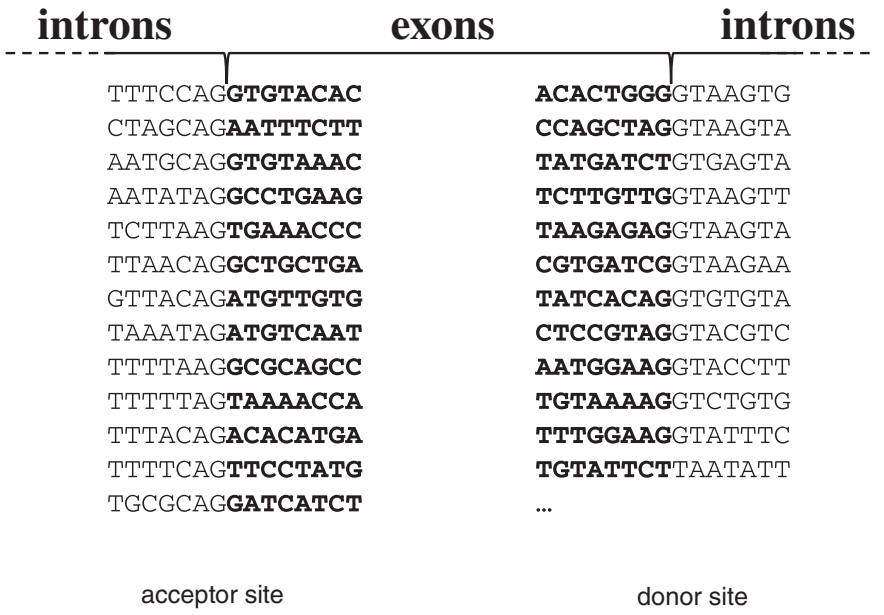


Fig. 5.4 *Splicing sites.* The figure represents the splicing sites found in the sequence of the gene encoding the mouse leptin receptor (*Lepr*-Chr 4). The intronic sequence of the donor (GT) and acceptor (AG) sites are highly preserved (this is known as the GT-AG rule)

forms, as opposed to the major transcript forms, are mostly not conserved across species and are probably of recent origin (Modrek and Lee 2003) (Fig. 5.5).

Alternative splicing can generate a large variety of proteins from the same DNA coding sequence by modifying the exonic contribution of the mature messenger RNAs. It is clear that the actual number of genes in a species has only a relative meaning, since the splicing machinery can tremendously increase genetic diversity. In this context, the number of exons is certainly much more informative for researchers than the number of genes. Alternative splicing is considered to be a very important mechanism for resolving the discrepancy between actual gene number and organismal complexity.

The mechanisms regulating alternative splicing and leading to the incorporation, or lack thereof, of a specific exon into the final product (sometimes designated the *splicing code*) are not yet completely unraveled. These processes probably involve *trans*-acting proteins (repressors and activators) encoded elsewhere in the genome that pair with *cis*-acting regulatory targets on the pre-mRNA. It is also likely that the secondary structure of the pre-mRNA transcripts plays a role in the regulation of splicing (Barash et al. 2010).

As we will discuss in Chap. 7 and as demonstrated by hundreds of positional cloning experiments performed in the mouse and rat, splicing sites are common targets for the occurrence of mutations. These sites are not always in frame with the sub-modulation of the mRNAs in triplet, and can also occur within codons.

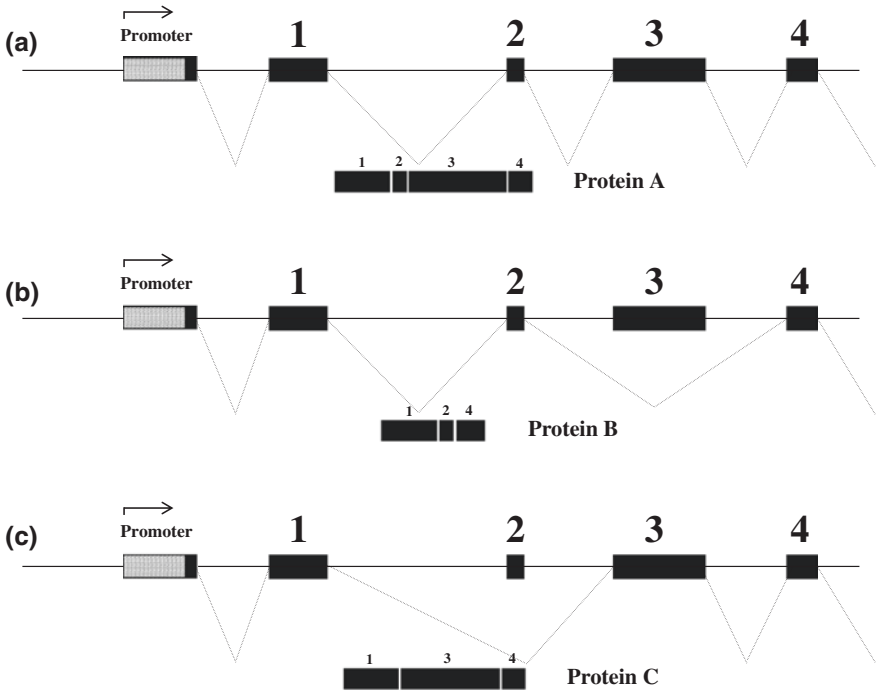


Fig. 5.5 *Alternative splicing.* In mammals, around 95 % of multi-exonic genes are alternatively spliced to produce different proteins (**A**, **B**, **C** …). Some exons are present in all transcripts (*constitutive* or *major forms* of transcripts), while others are *optional* or *alternative*. Exons of the second type are mostly not conserved across species and are probably of recent origin. For orthologous genes, the number of exons is sometimes variable among species, and the presence of recently captured exons is sometimes observed

In these cases, of course, the two contiguous exons are inseparable and are jointly incorporated into the transcript or skipped. The mRNA transcript, once adequately spliced, receives a cap of a methylated guanine nucleotide that is added to its 5' end to protect it.

The enormous amount of information collected by mouse geneticists indicates that the average size of a mouse gene is approximately 30–40 kb at the DNA level, while the average mature or processed mRNA molecule (mRNA mature transcript) is approximately 2 kb. The average gene density is in the range of 1 gene per 95 kb of DNA, i.e., very close to the predictions. The smallest (known) gene is 0.1 kb and encodes the t-RNATyr. The largest gene is *Titin* (*Ttn*-Chr 2), with 2.8 Mb of genomic sequence and 363 exons producing a spliced mRNA larger than 100 kb. The introns are also of various sizes, ranging from around 0.5 kb for the short ones to 30 kb for the longest (dystrophin-*Dmd*), with an average intron size of 4.7 kb. For the exons, the shortest consists of only 9 bp (exon 29 of *Myo5a*), and the largest is 7.6 kb long (exon 26 of *Apob*), with an average exon size of approximately 290 bp. Altogether, when added up, the exons represent

1.2 % of the total mouse DNA, the introns 26.7 %, and the intergenic regions 69.3 %. The number of exons per multi-exon gene varies from 1 to 363 with an average of 8.4. Finally, around 4,000 genes have only one exon.

The configuration of the "typical" mouse structural gene, as we just outlined it, is probably very similar to the average mammalian gene, and this is a blessing for the establishment of comparative maps; in short, the DNA sequence of two (not only one) mammalian genomes is an invaluable tool for making predictions about a third one. Many examples could be obtained from the cross-comparisons of mouse, rat, and human sequences. As of today, 17,054 mouse genes have an orthologous copy in the human genome, while 18,458 mouse genes have an orthologous copy in the rat. Finally, a total of 20,388 mouse genes have orthology annotations with at least one other species.

The classical gene we just described corresponds to a protein-coding gene. In fact, we now know that this category of genes represents only a proportion of the genes in the mouse genome that specialists consider to be in the range of 25–30 %. Most other genes encode RNA molecules of various sizes: some have an *open reading frame* (ORF) but most do not. Some are spliced, others are not, and the majority of these transcripts are processed further in smaller molecules. Most of the RNAs stay in the nucleus, suggesting that they have a function. Finally, all these RNAs exhibit a rather low degree of interspecific homology, indicating that the selection pressure they experience is of a different type. We will discuss this point more extensively at the end of this chapter when discussing the mouse transcriptome.

In November 2014, the Mouse Genome Informatics database estimated the number of mouse genes with nucleotide sequence data at 34,628 and the number of genes with protein sequence data at 24,553. This information seems reliable when compared with other species. Out of these genes, only 16,345 have experimentally based functional annotation.

Finally, we must point out that the distribution of genes in the mouse genome is very uneven. Mouse chromosome 11, for example, has twice the gene density of chromosomes 10 or 12, and the Y chromosome has only a few genes in an "ocean" of repeated DNA.

5.3.3 Finding the Regulatory Sequences

One of the biggest challenges of genome annotation is to identify gene regulatory regions. These comprise proximal and distal regulatory elements, according to their distance from the transcription starting point. Proximally are the *promoters* and associated promoter elements. Distal elements are enhancers, silencers, insulators and locus control regions (Fig. 5.3). Proximal and distal elements are usually composed of clusters of short intermingled transcription factor-binding DNA motifs referred to as modules or cis-regulatory modules (CRMs) (Hardison and Taylor 2012).

DNA sequence and local chromatin landscape act jointly to determine transcription factor (TF) binding intensity profiles. As a result, a regulatory module is defined by its sequence, since it binds transcription factors, and is thus expected to contain specific binding sites for these. It is further defined by its accessibility to TFs, which is linked to chromatin structural specificities such as histone modifications and local occupancy by nucleosomes. These are highly dynamic events which reflect the history of the cell and which are responsible for differential gene expression in animal development and cell differentiation. This implies that canonical binding sites for transcription factors are seldom sufficient to define a regulatory module, and methods relying on binding site identification usually have a high rate of false positives. For example, out of 132 regulatory modules predicted by algorithm analysis to bind TCF4 (a key transcription factor in the WNT1 signaling pathway), only 10 were validated using chromatin immunoprecipitation (ChIP)—little more than a random representation (Hatzis et al. 2008). This further implies that most CRMs will be difficult to identify until the chromatin landscape around them is defined. As a result, whereas the transcriptional apparatus reads the regulatory elements in the genome very efficiently, we still lack a universal syntax to decipher them, and this is quite critical: for regions that are defined by an unequivocal syntax, such as the coding exons, mutations can be characterized by just sequencing the whole mutated genome, together with low-resolution meiotic mapping, using no more than two dozen F2 mice (Xia et al. 2010; Arnold et al. 2011). Reaching the same level of power for regulatory regions would change the face of gene regulation analysis. Fortunately, this field is developing at a rapid pace, following the systematic reliance on strategies that directly measure sequence occupancy by Transcriptional Regulatory Factors (TRFs) within the living cell, such as chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) or DNase I digital genomic footprinting, which are currently performed or compiled by ENCODE (the ENCODE Project Consortium 2011—see above). Most of these results to date have been obtained for human but major conclusions also apply to the mouse, as demonstrated by results already obtained in this species.

Proximal regulatory modules (PRMs) at and around transcriptional start sites (TSSs) are the most straightforward regions to identify, since the TSS is accessible from the transcription product, the RNA. Cap-analysis of gene expression (CAGE) and RNA sequencing (RNA-seq) have contributed to the definition of TSS and consequently of PRMs. From these analyses, it appears that mammalian promoters can be separated into two classes: evolutionarily conserved promoters bearing a TATA box, and more plastic, evolvable CpG-rich promoters. The latter are by far the most frequent promoters since the TATA box (with a core DNA sequence 5'-TATAAA-3') is found in only one quarter of all promoters in a mammalian cell, usually around 30 bp upstream of the transcription start site. The TATA box, the first core promoter element identified in eukaryotic protein-coding genes (Goldberg 1979), is an anchoring site for the pre-initiation complex of transcription involving RNA polymerase II. The CpG sequence works similarly via the Sp1 factor. A CAT (or CCAAT, or CAAT) box, with a

consensus sequence GGCCAATCT, is inserted upstream of the TATA box, 75–80 bp from the transcription start site. Some genes with relatively ubiquitous expression do not have this GGCCAATCT sequence. In CpG-rich promoters, the start sites are usually multiple and organized in clusters at the 5' end of the gene, whereas TATA-box-bearing promoters have a single or at least a predominant start site. As of 2006, 729,504 potential mouse TSS sites were defined, organized in 159,075 clusters, a figure that far exceeds the number of genes identified (see above) (Carninci et al. 2006). Furthermore, mapping techniques such as CAGE are quantitative and provide a measure of the amount of transcription initiation in any given genomic region or for a given gene, in different tissues.

The situation is much less clear for distal regulatory elements such as silencers, enhancers or locus control regions. Enhancer elements, which can be located at some distance from the core promoter elements, where the transcription initiation apparatus is bound, are sites for fixation of transcription factors. The enhancer-bound transcription factors bind co-activators such as Mediator and p300, which in turn bind the transcription initiation apparatus, thus providing a link between enhancers and promoters. Non-coding RNAs may be associated with Mediator in this process (Lai et al. 2013).

Constraints on distal regulatory elements appear rather loose. Enhancers have been located at the 5' or 3' ends of coding regions, within introns and even within coding exons (Birnbaum et al. 2012), where they impose a further layer of constraint on the coding sequence. They can be close to the transcription start site or, in contrast, extremely remote (one to several megabases)—not to mention the possibility of them lying on a separate chromosome, from which they act in *trans* on a gene-coding region (Savarese and Grosschedl 2006). Furthermore, there is no evidence that the closest enhancer to a gene is the one likely to be active on this gene (Li et al. 2012). In cases when regulatory modules are remote, mutations that affect them may lie within another gene. For example, the CRM driving *Sonic hedgehog* (*Shh*) expression in the limb lies within the intron of another gene, *Lmbr1*, which for some time puzzled geneticists (Hill 2007). Similarly, a *Gremlin1* (*Grem1*) CRM lies within the *Formin* (*Fmn1*) gene, such that the latter was long considered as responsible for a limb defect (its original name was *Limb deformity*), whereas it does not have any known function in limb development, contrary to *Gremlin*.

These difficulties will be overcome when most regulatory regions have been defined according to transcription factor occupancy using strategies such as ChIP-seq. There is still a long way to go: according to experiment matrices recently published by UCSC Genome Bioinformatics, only 13 out of about 60 known histone modifications and 120 out of the estimated 1,700–1,900 transcription factors have been examined to date in the human genome by ChIP-seq. These, furthermore, have been analyzed in a number of cell lines in culture (which often bear little similarity to cells within organisms) or in readily accessible adult cells, such as blood cells, but many tissues in the adult, not to mention embryonic stages, have not been investigated—and the mouse genome lags

behind. Tissues, especially embryonic tissues, provide only sparse material, and methods will have to be miniaturized before they can be extensively analyzed. Nevertheless, the power of these new strategies is such that we can be confident that, in the near future, the regulatory syntax of the genome will be worked out. One major difficulty that may remain is attributing a given CRM to a specific gene in a defined physiological or developmental context, since, as we have seen, enhancers may be very remote and there is evidence that the closest enhancer to a gene is not necessarily active on that gene. Assessing the correlation of the chromatin state at enhancers and RNA-PolII occupancy at promoters, for each possible enhancer–promoter pair of elements in a chromosomal domain, may help define enhancer–promoter organization (Shen et al. 2012). This may be insufficient, due to the properties of enhancers discussed above. We see that the regulatory sequences of a gene can hardly be circumscribed a priori. At this stage, genetic approaches may prove very helpful, since, following mutagenesis, a phenotype attests to an alteration that affects one gene with no a priori hypothesis on the regulatory mechanisms for this gene. Unfortunately, ENU mutagenesis is much more efficient at mutating coding sequences than regulatory sites, for reasons that are not entirely clear. It may be because regulatory regions are often redundant (Lagha et al. 2012), and there may be multiple TRF binding sites within an enhancer, making it unlikely that a single mutagenesis experiment will abolish all the binding sites. In contrast, exceptions to this rule have proven highly educational. This is the case for the limb-specific regulatory module of *Shh*, which is located nearly 1 Mb (~0.6 cM) upstream of the coding region and has been extensively characterized via genetic strategies.

These strategies take advantage of several assets of genetic tools. First, they allow a fine mapping of the genetic alteration. This may be very valuable in the case of distant regulatory sequences. It should nonetheless be kept in mind that CRMs are often too remote for molecular walking strategies along the chromosome, but too close for genetic segregation and localization. While a huge number of polymorphisms have been defined in the mouse genome (SNPs), the precision of mapping still depends on the possibility of getting them to segregate in a cross—i.e., the number of meioses that can be analyzed (with 1,000 meioses yielding a 0.1 cM precision). In a historical attempt to localize *Hx*, a limb mutation that turned out to affect the distant CRM of *Shh*, analysis of more than 2,000 meioses in a cross involving *Mus m. castaneus* reduced the candidate region to a little more than 400,000 bp—a genetic tour-de-force, but still insufficient to identify a causative point mutation. At a minimum, genetic mapping based on segregation defines boundaries within which the regulatory sequence can be sought by other approaches.

To characterize the affected sequence in a mutant, an essential strategy is the reliance on multiple alleles for the mutation. It is even better to have alleles of a different nature in addition to point mutations (insertions, translocations), to allow easier entry points into the mouse genome. Thus, for the *Shh* CRM, as for *Gremlin 1* (*Limb deformity—Grem1*^{ld}, another limb mutation), a transgene

insertion provided an entry point to the CRM (Lettice et al. 2002). This illustrates the value of mutagenesis strategies that generate chromosomal accidents (deletions, translocations, transposon insertions—see Chap. 3) to locate regulatory modules. Examples include *PiggyBac*, *Sleeping beauty* or *Tol2* transposons, and ethyl methane sulfonate (EMS)-induced deletions in ES cells (Munroe and Schimenti 2009).

It has been shown over the past few years that many CRMs are active on more than one gene, defining so-called "regulatory landscapes". Thus, many genes in the landscape show the same expression profile as the gene of interest and may be suspected to encode proteins acting in *trans* as regulatory factors. Examples such as *Shh* (CRM within the *Lmbr1* gene) and *Grem1* (CRM within the *Fmn1* gene) are illustrative in this respect. In such cases, it is essential to define whether regulation occurs in *cis* or *trans*, and, up to very recently, only genetic tests could unambiguously settle the issue. The principle of the test is straightforward, but requires that two allelic forms of the regulatory region and its target, respectively, can be discriminated in a genetic cross. When the regulatory sequence is defined by a mutation, this provides the differential allele for the CRM. The gene acted on must have two alleles, either coming from different mouse subspecies or one being an engineered allele. The ultimate demonstration that the characterized alteration in the genome is the cause for the abnormal phenotype will be provided by recapitulating this phenotype using the altered sequence in a functional test, such as expression of a reporter in a transgenesis experiment, or phenotypic rescue by BAC transgenesis, or de novo creation of the suspected mutation by homologous recombination.

With its very powerful tools (different mutagenesis strategies to generate different types of mutations, screens to identify new dominant and recessive mutations, *cis–trans* tests, etc.), genetics could play a major role in the identification of new regulatory modules. However, genetics now has strong competitors over the whole spectrum: targeted mutations, long-range haplotyping by genome sequencing strategies, and identification of remote regulatory modules by scanning the genome via overlapping transgenes. Even before we can directly identify CRMs using appropriate algorithms, genetic approaches may be outdated by genomic strategies—which also are considerably less expensive.

5.3.4 Organization of Syntenic Regions at the Chromosome Level

As we explained in the previous chapters (Chap. 4 in particular), the linear arrangement of mouse genes along the chromosomes tends to be preserved, at least to some extent, among the different species of mammals, recalling the existence of a more or less distantly related common ancestor. This means that when two genes are found closely linked in the mouse, they have a good chance of also being linked in the rat and in human genomes, depending on the degree

of linkage. With the ever-increasing resolution of genetic maps and the availability of genomic sequences of several different mammalian species, it has become possible to reconstruct the progressive reshuffling of the chromosomal segments that occurred across the species in question during evolution. For example, scanning the human, mouse, and rat genomes at high resolution we find that there are 280 orthologous chromosomal segments between human and mouse, 278 between human and rat and 105 between rat and mouse. Comparisons between dog, cat, and cow, whose genomes are also completely sequenced, indicate that the number of chromosome breaks between human and rodents (~280) is consistent with the number of synteny breaks observed in other species separated by similar evolutionary distances. However, the number of chromosomal rearrangements between rat and mouse seems to be excessive if the divergence between the two species really occurred 12–14 Myr ago. Explanations for this discrepancy are lacking.

The existence of these homologies of synteny indicates that, during evolution, many genomic segments of the different species have been broken and then translocated, inverted, or transposed several times. This, however, is difficult to reconcile with the experimental observations presented earlier, indicating that most alterations in the karyotype structure are in general strongly counterselected by impeding normal gametogenesis in heterozygotes. Here again, explanations are awaited to reconcile all these observations, but it is tempting to speculate that this may be linked to the mechanisms of speciation themselves.

Homologous chromosomal segments display great variations in size across the different species. Mouse chromosome 11, for example, contains a large homologous region (almost all) to human chromosome 17q, while some other homologous chromosomal regions are extremely small-sized, and are sometimes reduced to a few genes. Human chromosome 21 has homologies with at least three mouse chromosomes (10, 16, and 17) and this, as we already mentioned, has hampered the development of mouse models of Down syndrome.

When checked at high resolution, it is sometimes observed that the genes in one species are not exactly in the same order as in another related species, although they are within the same syntenic segment. The genes flanking the *OAS* cluster on human chromosome 12q are on the same syntenic segment as the orthologous genes on mouse Chr 5, but are not in the same order, because a short inversion occurred in one of the lineages (probably in the mouse). Many other such rearrangements have been observed in other regions of the genome (Fig. 5.6).

Based on observations made in several distantly related eukaryotic species, the hypothesis has been suggested (Petkov et al. 2007) that the associations or clustering of genes within short genetic distances might have occurred initially because the genes in question were cooperating in various cellular and physiological functions (akin to large operons, so to speak). It is then not so surprising that these associations have remained relatively unchanged during evolution. Some support for this interesting hypothesis has been provided by the observation of non-allelic parental associations in recombinant inbred strains. Another stronger line of support should come from the analysis of the genome sequence of mice from the Collaborative Cross (see Chaps. 9 and 10).

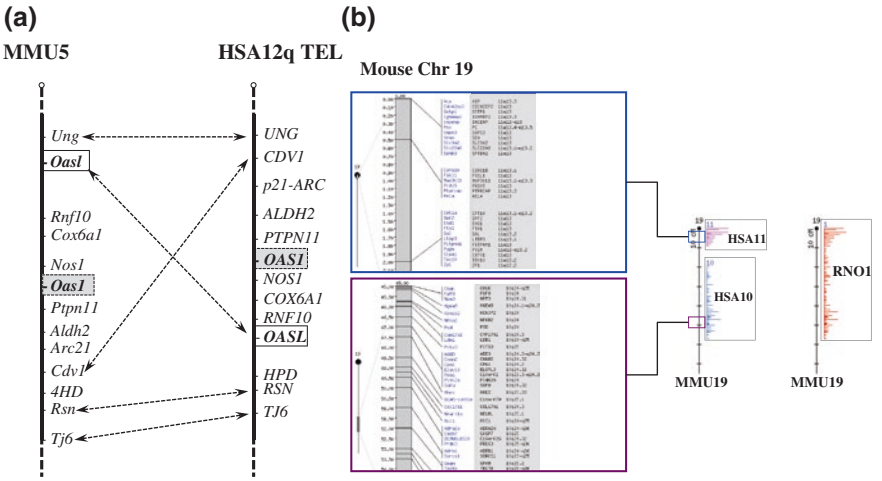


Fig. 5.6 *Homologies of synteny.* **a** An example of homology of synteny between mouse Chr 5 and human Chr 12q24 in the region of the *Oas/OAS* cluster. The genes flanking the *OAS* cluster on human Chr 12q are on the same syntenic segment as the orthologous genes on mouse Chr 5, but not in the same order because a short inversion occurred in the mouse. Many rearrangements of this kind have been observed in other regions of the genome. **b** Another example of homology of synteny between mouse Chr 19 and human Chrs 10 and 11. The same mouse Chr 19 also exhibits homology of synteny with a large fragment of rat Chr 1. More than 90 % of mouse and rat genes are in regions exhibiting homology of synteny with a chromosomal region in humans (the maps are from MGI database—2013)

5.3.5 Gene Families and Pseudogenes

As we already mentioned, when looking in the mouse genome for a DNA sequence orthologous to a human or rat gene we generally find them in the homologous syntenic region, as expected. However, it is not uncommon to find that the sequence homology between the two species is not always in a 1:1 ratio. On human chromosome 12q, for example, there is a cluster of three genes encoding 2′,5'-oligoadenylate synthetase (OAS), an enzyme that is induced by interferons and plays an important role in the inhibition of cellular protein synthesis and resistance towards viral infections. In this cluster, the human genes are arranged in the following order: HSA12 *cen—… –OAS1–OAS3–OAS2– …– tel.*

When looking for the orthologous syntenic region encompassing the OAS encoding genes in the mouse genome, we find a cluster on chromosome 5 with no less than ten genes. These genes exhibit a very high degree of sequence similarity and the linear order: MMU5 *Cen–… –Oas2–Oas3–Oas1e–Oas1c–Oas1b–Oas1f–Oas1h–Oas1g–Oas1a–Oas1d– …– Tel*. Thus, the human *OAS2* and *OAS3* genes each have, and as expected, a single 1:1 orthologous copy on mouse chromosome 5 while the human *OAS1* has no less than eight copies (1:8 orthologs). These *Oas1*

genes are all transcribed although not always in the same direction, indicating that they probably result from a series of segmental duplications with subsequent rearrangements (inversions). In the rat, the structure and organization of the cluster is similar to that of the mouse, but with only eight genes; the orthologous copies of mouse *Oas1a* and *Oas1e* are missing (Perelygin et al. 2006). These differences between the human, rat, and mouse OAS clusters indicate that the genomes of these three species are in constant evolution. Similar observations have been made when performing sequence alignments between mice of the same genus *Mus* but belonging to different species (Fig. 5.7).

These clusters of genes (the three human genes, ten mouse genes and eight rat genes), which encode proteins with similar biochemical functions, were presumably formed by recurrent duplications of a single ancestral gene and represent what geneticists call a *gene family.* Such gene families are common in mammalian genomes and include, for example, the genes encoding the globins, the myosins, the *Hox* and *Sox* clusters, etc. Looking at different unrelated vertebrate species, one observes that the number of repeated copies is highly variable, and the significance of these variations in copy number (if any) is not clear. In the case of the mouse *Oas* cluster, all ten copies are transcribed but the mouse *Oas1b* gene carries a stop codon in its exon 4, resulting in the premature truncation of the

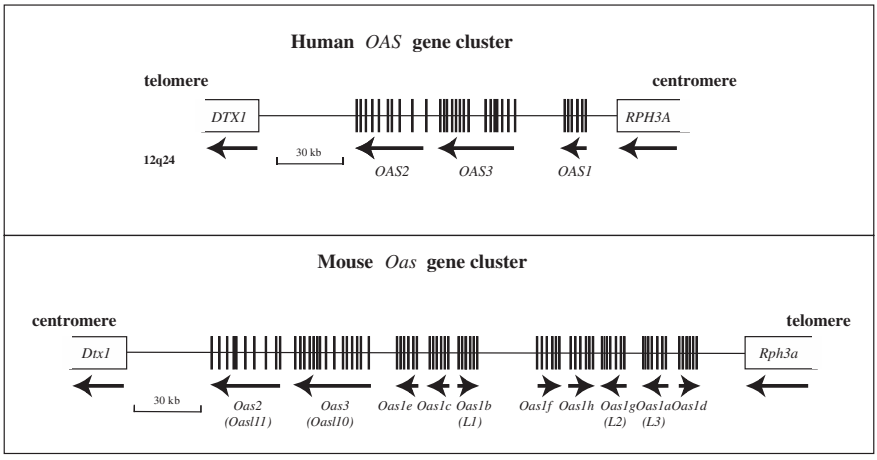


Fig. 5.7 *Gene families.* The three genes encoding human 2′, 5' oligoadenylate synthetase (OAS) are clustered on HSA12, flanked by the same two genes (*DTX1* and *RPH3A*) as in the mouse, and ordered as indicated in the figure. These three genes are transcribed in the same direction. The homologous region is on mouse Chr 5 (MMU5) and consists of ten genes with a very high degree of sequence similarity. The orthologous copies of human *OAS2* and *OAS3* are well preserved, with a 1:1 orthology, while human *OAS1* has no less than eight orthologous copies in the mouse. This cluster of *Oas1* genes probably results from a series of segmental duplication with subsequent rearrangements (inversions). All these genes are transcribed, although not always is the same direction. Such quantitative differences between the human and mouse OAS clusters indicate that the genomes of these species are in constant evolution, although with variations in gene copy numbers (Adapted from Mashimo et al. 2003)

gene product (oligoadenylate synthetase or 2′,5'-OAS), leading itself to its complete inactivation in virtually all mouse laboratory strains. Interestingly, this mutation does not exist in wild mice and researchers demonstrated that this difference, which is specific to the *Oas1b* gene, is responsible for the susceptibility of practically all laboratory mice to experimental infections with flaviruses. The function(s) of the proteins encoded by the other genes of the family is (are) not yet elucidated but, obviously, they do not complement the functional deficiency of *Oas1b* in laboratory strains.

The formation of a gene family results from a mechanism that is classical in evolution. As in the case of the *OAS/Oas* clusters, a majority of these families are formed by a succession of tandem duplications of a single ancestral gene and the different proteins encoded by the genes of the same family (commonly designated *isoforms*) generally have similar biochemical functions, but this is not a rule. Some gene families are easy to identify because the duplicated copies are closely linked to each other, are arranged in tandem, and have retained similar sequences. In other instances the situation is more complex because the gene family is ancient and has been more or less extensively remodeled during evolution. This is the case, for example, with the *Oas1* gene cluster that we described above and two other genes with a similar structure (Oas-like1 and Oas-like2—symbols *Oasl1* and *Oasl2*), located 4 cM away, on the proximal end of the same mouse Chr 5 (Fig. 5.8).

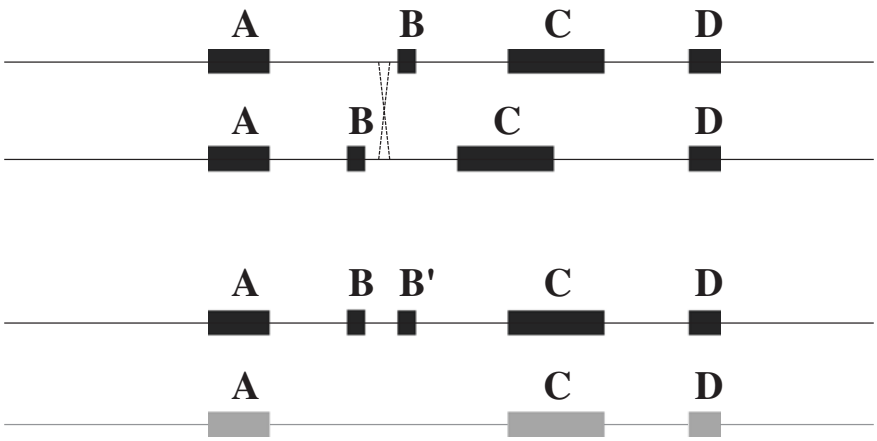


Fig. 5.8 *Gene duplications*. Some tandem duplications result from unequal crossing-over between homologous chromatids, as indicated in the illustration. The chromosome with a deleted gene or segment (in *grey*) is in general rapidly lost, while the duplicated region (*solid black*) is retained. Gene duplications can also result from error (slippage) of the polymerase during DNA replication, with the enzyme copying the same segment more than once. Gene duplication is an essential source of evolutionary innovation when the duplicated copies acquire specific functions (for example, the different isoforms of β-globin, *Hox* genes etc.)

This is also the case for the genes encoding the globin subunits, which are all clearly derived from a single ancestral copy that existed some 500 Myr ago, but are now separated in two different clusters in the mouse genome (α-globin on Chr 11 and β-globin loci on Chr 7). The expansion or contraction of gene families in a specific lineage can be due to chance, or can be the result of natural selection, and it is extremely difficult to decide between these two options.

When genes are duplicated in tandem, it is also common to observe that not all the copies are transcribed in exactly the same way. For example, according to the strain, laboratory mice have either one or two copies of the gene encoding *Renin*, a protein that participates in the regulation of arterial blood pressure (*Ren1* and, sometimes, *Ren2*-Chr 1). *Ren1* encodes the renin mRNAs found in the submaxillary gland while *Ren2* encodes the renin mRNAs found in the kidney. This difference in transcriptional activity can be explained by the promoter regions of these two genes, where structural differences have been described (Panthier et al. 1984).

Some specific gene families, like those concerned with a reproductive function (exhibiting, for example, spermatid or oocyte-specific expression), an immunological function, or an olfactory function (encoding, for example, the odorant (OR) or vomeronasal (VR) receptors) originated from relatively recent duplications (expansions) that occurred in the mouse lineage since the time of its divergence from the rat, around 12–14 Myr ago. In the initial draft of the C57BL/6 genomic sequence, for example, scientists were surprised at the identification of some 1,400 OR genes and 332 VR genes. In the human genome the same olfactory or vomeronasal receptors are much less numerous. The explanation generally proposed to explain these considerable differences is that such sequences are preserved because they are translated into functional proteins that are more or less important for the host species. Geneticists have coined the expression "*genome shaping*" to account for such a situation where the genome structure is influenced by natural selection triggered itself by environmental factors (Nouvel 1994). Although one can accept the idea that olfactory receptors are much more important for wild mice than for human beings, the same argument is less obvious for some other genes that are members of very large gene families in rodents but are much less represented in the human genome.

After careful examination and comparison with a consensus (or ancestral) sequence, it is common to observe that some members of a gene family carry point mutations (SNPs). These mutations are missense or sometimes nonsense, resulting in a loss of function for the gene in question. This is the case for the Oas1-like gene (*Oasl1*) described above. When this occurs, the mutated gene no longer encodes a functional protein, even if it is still transcribed. It is then classified as a *pseudogene* and its sequence will progressively degenerate, generation after generation, until it becomes unrecognizable in terms of structure. The pseudogene is then called a *relic*, a *vestige* or a *fossil*, and the intergenic regions of the genome have sometimes been described as "cemeteries" for these degenerated genes. The "death" of a gene is not important for the survival of the species as long as other copies of the family are present in the genome as potential backups, capable of taking over the function of the missing copies.

When missense mutations (i.e. leading to an amino acid substitution) occur in a gene that is a member of a family, this results in the gene encoding a novel protein, with sometimes new characteristics, a different 3D shape, a different stability, etc. Evolution will then "decide" whether this novel protein deserves to be retained or not based on the potential advantages it may confer to the affected individual in its current environment (Demuth and Hahn 2009). In this case, one realizes that diploid organisms have an advantage since they can put to test, in the same genome and for a few generations, both the ancient and the new copy (allele) of a given gene and finally retain the one with the best fit.

An interesting gene family is that of myosins, mostly known for their role in muscle contraction but also involved in a wide range of motility processes. In fact, myosins belong to a huge superfamily of genes whose products share the basic properties of actin binding, ATP hydrolysis (ATPase enzyme activity), and force transduction. Virtually all eukaryotic cells contain myosin isoforms (alternative forms). Some isoforms have specialized functions in certain cell types (muscle), while others are ubiquitous.

A careful analysis of the initial draft of the mouse genome sequence indicated that, in this species, the rate of nucleotide substitution is approximately twice as fast as the rate in human, and this explains why, after a few million years, it is sometimes difficult to establish sequence similarities between some elements of the human and mouse genomes.

As we discussed, it is clear that the mammalian genome contains a great number of sequences that look like protein-coding genes but, in fact, are not (or no longer). The first category of these sequences is the *pseudogenes* we reported above, which are duplicated copies of an ancestral (single copy) gene, and have become non-functional after the accumulation of random mutations (SNPs or indels). There is, however, another category of pseudogenes that geneticists call *processed pseudogenes*. These pseudogenes, unlike the former ones (which are then called *unprocessed pseudogenes*), originate from the retrotranscription of messenger RNAs back into the genomic DNA in more or less random locations. They have no introns and often exhibit mutations in their sequences (including frame-shifts and stop codons), indicating that they definitely do not encode proteins. Around 18,000 such pseudogenes have been identified in the mouse genome assembly (build 38.1), but their identification is often difficult. To discriminate between a true, *bona fide* gene (a gene encoding a protein and then submitted to purifying selection) and a pseudogene (processed or unprocessed), researchers calculate the so-called K_a/K_s ratio. This ratio compares the number of non-synonymous substitutions (K_a) to the number of synonymous substitutions (K_s) in the sequence of the two genes. Synonymous mutations, as we will discuss later, do not modify the amino acid sequence (for example, the GGC codon becomes GGA but still codes for glycine) and accordingly can occur at the same frequency in genes and in the pseudogenes, with no consequence. Non-synonymous mutations, on the contrary, because they generally alter the protein structure, and often its function, are counterselected and are uncommon in functional genes. Computing the K_a/K_s ratio is then a reliable assessment of whether a gene is a "true gene" or a

pseudogene. K_a/K_s values approaching 1 are indicative of neutral evolution, suggesting a pseudogene. In addition, most mouse pseudogenes do not have an orthologous copy in the same syntenic position in the human or rat genomes, whereas active genes generally do.

As we discussed above, most pseudogenes were considered to be "fossils" or "relics" of genes that, once transcribed and reintegrated into the genome, became silent and functionally useless. This view, however, might not be correct or universally true. In fact, there has been speculation and some evidence has been collected suggesting that pseudogenes, or portions of the latter, may be transcribed from the opposite strand relative to their functional counterparts, making them a source of antisense RNA. These RNAs have been proposed to play a role in the fine regulation of genes of the same family through RNA–RNA interaction (Balakirev and Ayala 2003). Even more recently, scientists working on the mouse transcriptome have identified no less than 10,000 full-length cDNAs derived from expressed pseudogenes—representing approximately 10 % of the known transcriptome—with a good half of them likely participating in various regulatory mechanisms. As noted by the members of the FANTOM 3 project (see later in this chapter), we must remain open-minded about the potential function of expressed pseudogenes. For this reason, pseudogenes have been referred to as "*potogenes*" (potential genes) (Balakirev and Ayala 2003; Hayashizaki and Carninci 2006).

5.3.6 Copy Number Variations

In a preceding section (see 5.3.1.1), while discussing the different structural variations that have been observed at the genome level, we noted that some genes have been found to be missing (deleted) in some mouse strains and not in others (for example, *Snca* on Chr 6), while other genes, in contrast, were duplicated in some strains and not in others (for example, *Ren1* and *Ren2* on Chr 1). Variations of this kind are common in mammals and one can certainly expect many similar cases to be reported in the future, for example when comparing distantly related strains or sub-species of the same *Mus* genus. Many of these duplications are lost after a few generations, but a few of them may be retained, eventually after a few changes, either by chance or because they have an adaptive value. We have already discussed this point.

Copy number variations (CNVs) originate from both coding and non-coding regions of the genome. The mechanisms leading to these CNVs in a specific chromosomal region have not yet been completely elucidated, but it makes sense to consider a priori that CNVs are of three kinds. A substantial proportion probably results from unequal crossovers, producing both deleted and complementary duplicated genomic segments. Given that these chromosomal rearrangements often concern large segments, the duplications have a greater chance to be transmitted to the next generation than the deletions, which are generally unviable and lost.

Another type of CNV probably results from defects occurring during DNA replication (for example, defects in replication fork maintenance). This class of CNV commonly occurs in somatic cell lineages (especially in neoplastic tissues), and, accordingly, occurs independently of the process of meiotic recombination.

Finally, the observation that some short-length chromosomal duplications have been found on different chromosomes (cases have been reported in the mouse) suggests that these duplications are, in fact, transpositions of DNA segments very similar to those described earlier and classified as transcriptionally active pseudogenes.

In the mouse, around 100 well-dispersed regions across the 19 autosomes and the X chromosome have been shown to harbor CNVs. Their greatest preponderance is on chromosomes 7, 12, 14, and X, where some of them appear as large blocks.

The sequence homology between the different copies is >94 % on the average, and their size ranges from 62 bp to 8.6 Mb (with an average length of 250 kb). In total, if we include both the deletions and the duplications, this represents close to 10 % of all polymorphisms (excluding microsatellites), with short deletions being more frequent than insertions (Cutler and Kassner 2008).

CNVs involving large or very large chromosomal segments, although rare, have been observed by cytogeneticists using the classical techniques of fluorescence in situ hybridization. Nowadays, more sensitive techniques, like high-resolution comparative genomic hybridization (HR-CGH) or representational oligonucleotide microarray analysis (ROMA), are adapted to this sort of analysis. Using appropriate DNA arrays, these techniques allow for the detection of structural variations at a resolution of 200 bp (Egan et al. 2007) (Fig. 5.9).

In the near future, taking advantage of the recent advances in DNA sequencing technology, it should be possible to identify and quantify many more CNVs at high resolution in both human and mouse, allowing comparisons to be made at the individual level.

The occurrence of CNVs at the genome level translates to variations in gene dosage within the duplicated or deleted regions (0/1–1/1–2/1, etc.), and it makes sense to think that this may be causative or associated with some pathologies. A trisomic mouse, for example, can be regarded as carrying a single large CNV, since the only difference relative to a normal karyotype is an extra chromosome. This difference can nevertheless result in a severe and often lethal syndrome. A good example where a CNV has been found to be causative of a pathological syndrome is Charcot–Marie–Tooth, type A (CMT1A) disease in humans. This neuropathy was found to segregate with a ~1.4 Mb duplication on human chromosome 17p12 among the members of the same family, suggesting a possible causal relationship. Shortly after this observation, the gene coding for peripheral myelin protein 22 (*PMP22*), a component of myelin, was identified within the duplicated region and mutations in this gene were found to be also responsible for a clinical form of the disease very similar to the form associated with the duplication (Valentijn et al. 1992a, b). Finally, an almost perfect mouse model of CMT1A was created by pronuclear injection of a YAC containing a normal, intact copy of the

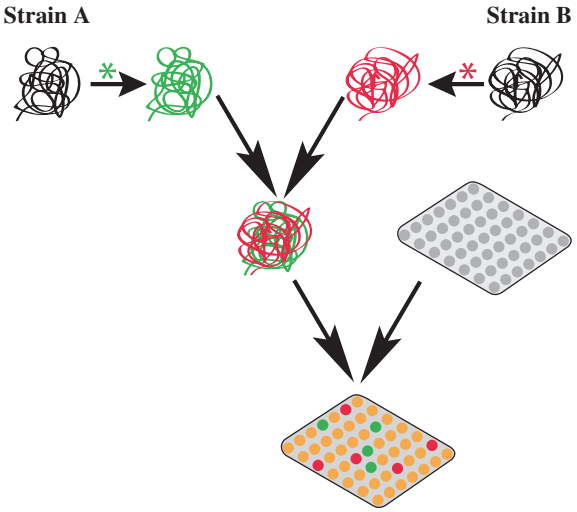


Fig. 5.9 *High-resolution-comparative genomic hybridization (HR-CGH).* This technique is useful for comparing two genomes, or two chromosomal regions, for possible quantitative differences in terms of copy number. The technique consists of two steps. First, a reference DNA sample is labeled with a fluorophore (for example, Cyanine 3, *green*) while the DNA from a test sample is labeled with a different fluorophore (for example, Cyanine 5, *red*). In the second step, equal quantities of the two-labeled DNA samples are mixed and co-hybridized to a DNA microarray of several thousand evenly spaced cloned DNA fragments previously spotted on the array. Finally, after hybridization, digital imaging systems are used to capture and quantify the relative fluorescence intensities of each of the hybridized fluorophores. Obviously, the ratio of the fluorescence intensities is proportional to the ratio of the copy numbers of DNA sequences in the test and reference DNA samples. If the intensities of the fluorophores are equal for a given probe, the spot appears yellow and the region of the genome is interpreted as having an equal quantity of DNA in the test and reference samples (i.e., no copy number variation (CNV)). If there is an altered Cyanine 3:Cyanine 5 ratio, this indicates a loss or a gain of the test DNA sample at that specific genomic region. Discovering which regions of the genome have undergone CNVs is achieved by another test, for example by sequencing followed by fine localization of the sequence. Finely estimating the CNVs can ultimately help to identify genes that are over- or under-expressed, or even deleted. The technique can be adapted to the localization of CNVs directly on the chromosomes

human *PMP22* gene and a large proportion of its flanking region. The conclusions of all these observations and experiments are that both point mutations and duplication of the *PMP22* gene can produce the same phenotype of severe demyelination in the peripheral nervous system.

If the mere duplication of an intact, normal myelin-encoding gene (*PMP22-Pmp22*) can induce a pathology in humans and mice, as demonstrated with YAC transgenics, one can then seriously consider that other CNVs might be at the origin of (or associated with) some clinical diseases or, at least, influence their phenotypic expression (penetrance or expressivity, for example) by altering the transcript level of some essential genes. The presence of some specific CNVs in the human genome has been found to be associated with susceptibility to autism

(Sebat et al. 2007; Cook and Scherer 2008). A reduction in CNVs involving the gene Defensin beta 1 (*DEFB*) has been reported to increase the risk of developing Crohn disease (Roberts et al. 2012). Other human pathologies are equally suspected to be associated with (or the consequence of) CNVs (e.g., autoimmunity, susceptibility or resistance to infectious disease).

In the mouse, genes involved in the control of the immune response or environmental sensory perception have also been found to exist in variable copy numbers in the genomes of the various inbred strains (Watkins-Chow and Pavan 2008). In these conditions, it should not be so surprising to observe in the future that these mice exhibit different phenotypes related to these CNVs.

Nowadays, many geneticists consider that the transmission of some complex traits might be better explained by the transmission of CNVs than by hypothetical Mendelian characteristics (Canales and Walz 2011). Observations relative to some infectious diseases in human populations have already provided preliminary clues to this important question. For example, Gonzalez and colleagues (Gonzalez et al. 2005) reported a strong positive correlation between a high number of copies of the gene encoding the chemokine CCL3L1 and HIV susceptibility.

5.3.7 Single Nucleotide Polymorphisms

When orthologous sequences from different mice (laboratory mice or wild mice) are aligned, single nucleotide differences are frequently observed in the DNA sequence. These differences are base-pair substitutions in most instances, less frequently insertions or deletions of one nucleotide. These sequence differences have been collectively designated *single nucleotide polymorphisms* (SNPs, pronounced "snips") and are the most common type of genetic variation at the DNA level. They are found in both coding and non-coding regions and almost all these SNPs are bi-allelic, i.e., presenting one of two possible nucleotides in an individual (e.g., homozygous G/G or T/T or sometimes heterozygous G/T).

SNPs are extremely abundant among the different mouse inbred strains, and even more so across the different strains recently derived from wild populations. These SNPs are easy to score and permit the performance of high-density/high-resolution mapping. They have undoubtedly been an important outcome of the mouse genome sequencing project, because they represent the ultimate genetic markers. We described their use and advantages in Chap. 4 (Fig. 5.10).

5.3.8 Tandem Repeated Sequences

Like other mammalian genomes, the mouse genome contains a large number of repeated (both coding and noncoding) sequences. They are classified as moderately or highly repeated sequences, and among the latter one must also

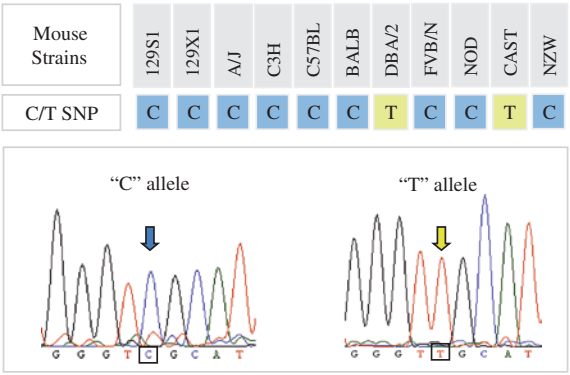


Fig. 5.10 *Single nucleotide polymorphisms (SNPs).* SNPs are single base-pair differences in the DNA sequence, and are the most common type of genetic variation. As described in Chap. 4, they are very useful for genetic mapping, they are found in both coding and non-coding regions, and almost all these SNPs are bi-allelic, i.e., presenting one of two possible nucleotides (e.g., G/G, T/T, or G/T genotypes). In the figure, the upper panel represents a C/T SNP that is polymorphic between DBA/2 and CAST (homozygous for the T allele) and other strains (homozygous for the C allele). The *lower panel* presents DNA sequencing electropherograms showing the SNP (*arrow*)

distinguish those that are organized as tandem repeats and those that are interspersed. *Tandem repeats* are those where the nucleotides motifs are repeated adjacent to each other in a head-to-tail manner. Depending on the number of nucleotides and on the size of the motif, these tandem repeats are known as *satellite* DNA (between 120 and 250 nucleotides), *minisatellites* (between 10 and 60 nucleotides), and *microsatellites* (between 2 and 6 nucleotides). In these types of repeats, the polymorphism is a direct consequence of the number of repeats. The interspersed or dispersed repeats are a totally different category and will be described below.

5.3.8.1 Satellite DNA

The name "satellite DNA" was coined in reference to a difference in the buoyant density of this category of DNA when compared to the density of bulk DNA. Satellite DNA constitutes about 5 % of total mouse DNA and is divided into two major categories: major satellite, which is composed of 234-bp repeats (6 Mb long altogether—occurring at a few loci on the genome), and minor satellite, which is composed of 123-bp repeats (from 500 kb to 1.2 Mb in size and located essentially in the centromeric and telomeric regions of chromosomes). Satellite DNA is the main component of heterochromatin, is not transcribed, and has proved to be rather difficult to sequence.

5.3.8.2 Minisatellites

Minisatellite loci are also known as *variable number of tandem repeats* or VNTRs. They consist of a short series of 10–60 bp repeated in tandem over and over to reach around 5–10 kb in size. They are extremely abundant and are distributed at more than 1,000 locations in mammalian genomes. The occasional slippage occurring during replication is probably at the origin of the minisatellite copy number variations, thereby making each individual unique (Kuznetsova et al. 2005). These highly polymorphic loci were used as genetic markers in the late 1980s, particularly in human studies, and became the basis for the famous DNA fingerprinting method that revolutionized forensic medicine. These "fingerprints" are the individual-specific band patterns resulting from the hybridization (by use of Southern blotting) of restriction-endonuclease-digested DNA with probes directed against extremely polymorphic minisatellite (VNTR) loci. Although it was used in a few mouse linkage studies and also for the genetic monitoring of inbred strains (isogenic individuals within an inbred strain share the same band pattern), the use of DNA fingerprinting in the mouse was abandoned after the advent of microsatellites as universal molecular markers (Julier et al. 1990; Silver 1995).

5.3.8.3 Microsatellites

Microsatellites (also known as short tandem repeats (STRs) or simple sequence length polymorphisms (SSLPs)) are tandem repeats of 1–5-bp elements that are probably the consequence of polymerase slippages. They are very abundant (approximately 10^5 copies per genome), extremely polymorphic, and widely distributed throughout the genome. Since the early 1990s, microsatellites have been the genetic marker of choice in mouse genetics because their analysis is extremely simple, inexpensive, and relatively reliable. For the same reason as for the SNPs mentioned above, we will review their interest as genetic markers in several chapters of this book and in various contexts (Fig. 5.11).

5.3.8.4 Trinucleotide Repeat Expansions

Some severe human genetic disorders have been found to be the consequence of the continuous and abnormal expansion of DNA-trinucleotide repeats in certain genes. The fragile X syndrome is one of these disorders and the first to be explained at the molecular level. Human geneticists found 230–4,000 CGG tandem repeats in a specific X-linked gene in affected patients compared with the 5–54 repeats in unaffected individuals. Similarly, Huntington disease (HD), which affects muscle coordination often associated with psychiatric problems, is caused

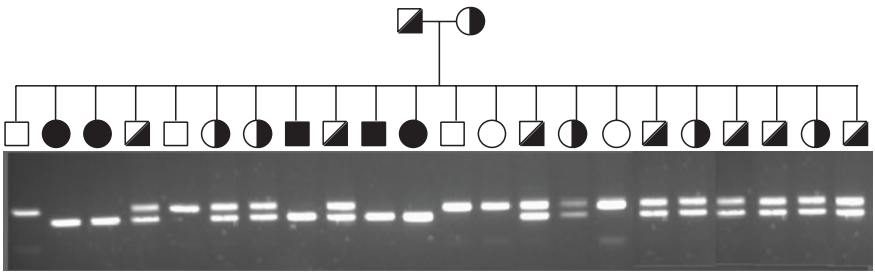


Fig. 5.11 *Microsatellites*. Microsatellites (SSLPs) are composed of short DNA sequences, measuring 1–6 bp, which are repeated in tandem a number of times. They are common in all mammalian genomes, where they exhibit variations in terms of the number of repeats (size polymorphism) and for this reason they are sometimes designated as simple sequence length polymorphisms (SSLPs). Microsatellites can be amplified by PCR with primers designed from the flanking regions. The number of repeats, which translates into size variations of the amplification product, can then be used as a reliable genetic marker. Microsatellites have been extensively used in the mouse for the establishment of high-density/high-resolution genetic maps and are still used for the acute localization of quantitative traits. As indicated in the figure, microsatellites are co-dominant markers, allowing for the identification of heterozygous genotypes. In the figure, we can observe the segregation of the alleles from a microsatellite locus on a pedigree. The male (*square*) and the female (*circle*) from this breeding pair are both heterozygous (*black* and *white*). In the 22 offspring we can clearly see the segregation of the two alleles, where some mice are homozygous (*solid color* on the pedigree) for one allele or the other (one band on the gel), and the rest are heterozygous (two bands). Note that the percentages are in agreement with Mendelian ratios for this co-dominant SSLP marker (~54 % heterozygous; ~23 % homozygous larger allele; ~23 % homozygous smaller allele)

by a CAG repeat expansion in the protein-coding regions of another specific gene called *Huntingtin* (*HTT*—in human Chr 4p16.3). In some other instances, such repeats are also observed and associated with a severe pathology but they are located outside of the protein-coding regions of the genes. To date, similar DNA-trinucleotide repeat expansions have not been reported in the mouse, but transgenic mouse models have been created by pronuclear microinjection of DNAs cloned from affected human patients (Ehrnhoefer et al. 2009).

5.3.9 Interspersed Repeated Sequences: Transposable Elements

Transposable elements (TE), as the name indicates, are small sized DNA sequences that move within the genome and insert into new chromosomal locations sometimes leaving behind a copy of their sequence at their original site (Wessler 2006). These TEs exist in virtually all genomes and have been described in bacteria, *Drosophila*, mammals and many other organisms. TEs were identified

and characterized for the first time in plants, more precisely in maize, through the somatic mutations they induced.[4] In the mouse, and more generally in mammals, these elements are repeated over and over, by thousands of copies, but they are dispersed in the genome and for this reason they are commonly designated *interspersed repeats* in opposition to the *tandem repeats* discussed above. Transposable elements are generally classified into two categories: (i) the *retrotransposons*, which transpose via an RNA intermediate in a "copy and paste" fashion, and (ii) the *transposons*, which use a "cut and paste" mechanism to move within the genome, with no RNA intermediate.

5.3.9.1 The Retrotransposons

Retrotransposons (or class I transposons) are of two kinds based on their size and structure: the LINEs (Long Interspersed Nuclear Elements) and the SINEs (Short Interspersed Nuclear Elements). In addition to these two kinds of transposons, endogenous retroviruses (ERVs) are often considered as equivalent to retrotransposons, as we will explain. Altogether these TEs represent the most abundant component of the mammalian genome, estimated at a proportion of greater than 40 % of genomic DNA.

Long Interspersed Nuclear Elements

The LINE family of retrotransposons, and more precisely the L1 subfamily, is the most important category of transposable elements in placental mammals, representing roughly 17–20 % of mouse genomic DNA. The normal, intact L1 sequence measures ~7.5 kb and consists of a promoter at its 5' end, followed by two non-overlapping open reading frames, ORF1 and ORF2, that encode respectively an RNA-binding protein and a 40-kDa protein with reverse transcriptase and endonuclease activity, and finally an AT-rich region of variable length at its 3' end. This basic structure is relatively uniform, but variations resulting from mutations or deletions, accumulated with time, are common. Thus, only a minority of the LINE elements (a few thousand) appears intact in the mouse genome. The mRNA transcribed from these LINEs serves as templates for the reverse transcriptase II encoded in ORF2, and this explains why this type of transposon is also designated *autonomous transposons*. The new cDNA (a new LINE element) is retrotransposed into a different site, at a new position in the genome, with the help of the endonuclease that nicks the chromosomal DNA and creates the conditions favorable for integration: in other words a true "copy and paste" mechanism. This

[4] Barbara McClintock was awarded the Nobel Prize in 1983 for the discovery of "jumping genes".

process of retrotranscription is similar to the one leading to the creation of processed pseudogenes, as discussed earlier. Sometimes it fails, and this also explains why so many LINEs are incomplete and truncated at their 5' end.

As observed after the sequencing of several mammalian genomes and comparisons between related species, L1 transposons are active as contributors to the so-called genome shaping and have been a source of evolutionary novelty by providing sequence motifs that can be recruited by the host, either for the regulation of its own genes or among its coding sequences. In contrast to this rather positive aspect, L1 transposition can also be deleterious for the host, for example when a transposed copy accidentally inserts within a gene or when it mediates a chromosomal rearrangement through ectopic (non-allelic) recombination (Sookdeo et al. 2013). The *spastic* mutation of the mouse (*Glrb*^{spa}-Chr 3), which is a model of human hereditary hyperekplexia (OMIM 149400), is caused by the intronic insertion of a 7.1-kb L1 element resulting in the aberrant splicing of the beta subunit of the glycine receptor mRNA (Mülhardt et al. 1994). L1 transposition can also be mutagenic in somatic tissues and was actually discovered through this type of activity in maize. This finding has potential consequences for the whole organism which can translate into an increase in cancer occurrences (Belancio et al. 2010). However, most L1 sequences are silenced by methylation and finally become inactive.

This mechanism of LINE retrotransposition, as described, would result in a progressive increase in the size of mammalian genomes unless a compensatory mechanism operates at some point. Based on recent observations, geneticists assume that the mechanism in question consists of repeated deletions (sometimes massive) of some of these constantly burgeoning sequences. Whatever the exact nature of the regulatory mechanism, the size difference observed between the human and mouse genome is generally attributed to variations in the number of L1 copies.

Short Interspersed Nuclear Elements

SINEs are a type of non-autonomous retrotransposon whose sequence does not encode any protein. SINEs have a sequence of around 100–500 bp, which is closely related to the sequence of some tRNAs or to short RNAs. The most common category of SINEs in the human genome is the *Alu*1 sequence, whose equivalents in the mouse genome are the B1 and B2 sequences. SINEs are transcribed by RNA polymerase III but their retrotranscription, necessary for their mobility inside the genome, is not completely elucidated and probably depends (at least in part) upon the LINE machinery—hence their occasional designation as non-autonomous retrotransposons.

There are around 1–1.5 $\times$ 10^6 copies of these SINEs in a mouse genome, representing between 11 and 17 % of the total genomic DNA. Depending on their sequence, SINEs are classified as lineage-specific (added to the mouse genome after the divergence from a common ancestor with other rodents) or ancestral

(before the divergence).[5] Thus, the sequences of these two categories of SINEs have great value for research in evolution and systematics.

Using a software program for multiple sequence alignment guided by phylogenetic trees, researchers have found a DNA sequence measuring 710 bp in the close vicinity of the bovine β-globin locus, sandwiched between two SINEs, and obviously resulting from a transposition (Zelnick et al. 1987). This finding may be considered circumstantial but it nevertheless indicates that, if such a transposition of a DNA segment (by "hitch-hiking", so to speak) can occur in the bovine genome it may also occur in other species, and this is important in the context of the constant remodeling of the genome structure.

The existence of a very large number of retrotransposons with nearly identical sequences, scattered throughout the mouse genome, has some potentially interesting technical applications in the sense that universal (non-specific) primers for PCR amplification can be designed based on the sequence of these retrotransposons and used either with another specific primer (for example, for cloning the sequences flanking a transgenic insertion) or with the same primer with the inverted sequence for the amplification of the host genomic DNA situated between two LINEs or SINEs.

The Endogenous Retroviruses

The *endogenous retroviruses* (ERVs) are a third kind of element that can affect the structure and function of the mouse genome. Although uncommon, infections of mouse germ cells by retroviruses can occur, resulting in the integration of more or less complete retroviral copies into the mouse genome. These retroviral copies are easily recognizable at the molecular level because they are flanked by two classical long terminal repeats (or LTRs) and contain the three classical genes *gag* (encoding structural elements of the virus), *pol* (encoding the reverse transcriptase), and *env* (encoding the coat protein of the virus). Many ERVs are incomplete and no longer move in the mouse genome, and in some cases one LTR is the only sequence that remains of an ancestral retroviral copy that has been completely excised or deleted.

Just like the LINEs and SINEs, ERVs occasionally have influence on the genome's structure and function. They can be mutagenic, like LINEs, when they integrate into the host DNA into or around a coding sequence. They can also trigger various forms of structural rearrangements. A classical example of the role of ERVs as mutagens is the *hairless* mutation of the mouse (Hr^{hr}) (Stoye et al. 1988; Cachon-Gonzalez et al. 1994). This recessive mutation is the result of the retroviral insertion of murine leukemia proviral sequences into intron 6 of a gene encoding a specific protein at the *Hr* locus of chromosome 14, which results in aberrant splicing of the gene. Many other mutations of this type have also been reported in the mouse. The viable yellow (A^{vy}) allele, which originated through the retrotransposition of an

[5] The ancestral SINEs are sometimes designated MIR3 (for mammalian-wide interspersed repeat elements).

intracisternal A-particle[6] (IAP) upstream to the canonical wild-type transcription start of the *agouti* gene (*A*), is another example.

Some elements of these ERVs can also have functional consequences. This is the case, for example, when long terminal repeats (LTRs) act as alternative promoters or enhancers leading to the transcription of tissue-specific RNAs. In humans, diseases have been reported as being caused by TE-generated alleles. These diseases include, for example, hemophilia A and B, severe combined immunodeficiency, porphyria, predisposition to cancer, and some cases of Duchenne muscular dystrophy.

Recombination between homologous retroviral sequences has also contributed to "gene shuffling" and to gene duplications and deletions that largely contribute to genome plasticity.

Several years ago, the retrotransposons we just described were considered as examples of the so-called "selfish" or "junk" DNA because, apparently, their only function was to make more copies of themselves with no apparent benefit for the host. Nowadays, the perspective has dramatically changed and these DNA elements are regarded as tools contributing to genome plasticity and "novelty". L1 sequences frequently insert into the introns of functional genes, where they can interfere with the transcription process without permanently harming the gene product. When the inserted L1 copy is long or very long, the transcription rate is reduced and this might represent another subtle (and reversible) method of gene regulation. When inserted into an intron, SINEs or LINEs can also introduce new splicing sites, allowing the de novo creation of new exons and accordingly of new protein domains. It is then up to the environment to determine, at no risk, whether the new protein presents some selective advantage, whether the structural alteration is selectively neutral or, on the contrary, whether it is detrimental and should be eliminated by returning to the original copy of the gene—which is still in the genome as a back-up. In other words, thanks to the TEs, evolution can perform experiments at virtually no cost.

5.3.9.2 The Transposons

Transposons exist in many species including bacteria, plants, insects (for example the P elements of *Drosophila melanogaster*), and mammals. They are relatively short elements, measuring a few kilobases when intact, and they encode an essential enzyme: a *transposase* (also called *transposonase*). The gene encoding this transposase is flanked by two inverted or palindromic terminal repeats that are essential for transposition in the genome. These terminal repeats pair with each other as the transposon folds and forms a loop. This DNA loop is then excised and released, ready to transpose into another location in the genome, hence the "cut and paste" mechanism of transposition.

[6] IAPs are a class of defective endogenous retroviral sequences measuring ~7 kb. These IAPs are mostly abundant in the endoplasmic reticulum.

The excision of a transposon from its original location in the host genome often generates a small gap in the genomic DNA, while its insertion in a new location disorganizes the neighboring genetic sequences. For these reasons the transposons are responsible for the occurrence of new mutations in the species where they are active.

In the mouse genome the vast majority of transposons no longer encode any functional transposase, and accordingly, they have lost the capacity to transpose: they are "dormant" or even "dead". Interestingly, a fish transposon, which had remained inactive for over 15 million years, could be artificially "resurrected" into an active one by the transgenic addition of two essential functional components into the same host genome: (i) the transposon DNA containing the two inverted terminal repeats, and (ii) the transposase enzyme essential for activation. This engineered (and resurrected) transposon, named *Sleeping Beauty* (Izsvák and Ivics 2005), has been shown to transpose efficiently enough in the mouse to be proposed as a tool for the in vivo production of mutations (Carlson and Largaespada 2005). This method of mutagenesis has the advantage that new mutations are created simply by breeding mice, and, most importantly, that the transposon DNA tags the integration site. However, the disadvantage is that the mutation rate is rather low, especially when compared to other mutagenesis methods. More recently, *SleepingBeauty* has also been reported as an interesting tool for cancer gene discovery and gene therapy (Copeland and Jenkins 2010; Howell 2012), helping for example to introduce transgenes into host genomes. Other resurrected transposons (*Piggy Bac* and *Mariner*) have also been used for the production of mutations (by gene trapping) and for transgenesis.

The transposable elements are definitely important elements of the genome, since they participate actively in its evolution. Together they are often referred to as elements of the "*mobilome,*" and it is likely that their role and functions are still underestimated.

5.4 The Transcriptome: Coding and Non-coding RNAs

In the same issue of the journal *Nature* announcing the initial draft of the mouse genome sequence (*Nature* 420–5 December 2002), another very important report was published, summarizing the results of the functional and manual annotation of a large collection (60,770) of full-length mouse cDNA[7] collected by the "FANTOM consortium" (Functional Annotation of the Mouse) of the RIKEN Genomic Science Center in (Okazaki et al. 2002). This publication, perhaps because it was released at the same time as the impressive and outstanding

[7] Full-length cDNA libraries are established from all RNA transcripts (protein-coding and non-protein-coding). Manual annotation of such libraries is a guarantee of their quality.

presentation of the mouse genome sequence, did not receive the attention we think it deserved from the community, at least when published. Ten years later, and based on the information gathered in the meantime from the analysis of the mouse and human genomes and transcriptomes, we think that this report should be considered another breakthrough in our understanding of the ways in which the mammalian genome actually works. Not only did it confirm some important observations that were made independently a few years earlier, for example about the unjustified overestimation of the number of protein-coding genes in the mouse genome (which was sometimes estimated to be as high as 120,000) and the concomitant underestimation (or mis-appreciation) of some other transcription products (Lander et al. 2001; Kapranov et al. 2002), but it also raised a number of new ideas that have been confirmed since and widely amplified in successive reports, in particular those of the same FANTOM consortium as well as in other reviews devoted to the analysis of the mouse transcriptome (Carninci et al. 2005; Katayama et al. 2005; Mattick and Makunin 2006; Gustincich et al. 2006; Saxena and Carninci 2011; ENCODE Project Consortium 2012; Kapranov and St Laurent 2012). The ideas that were developed in these initial reports have radically changed our views of the transcriptome, in particular the belief which was solidly anchored in most scientists' mind that proteins were the most important (if not the only) bioactive molecules encoded in the genome.

The main conclusions of the reports in question are the following: (i) the protein-coding RNAs (the mRNAs) and the other RNAs that cooperate with mRNAs in protein synthesis and processing (rRNAs, tRNAs, snoRNAs, and snRNAs) represent only a minor (around ~2–3 %) component of the transcriptome; (ii) the mouse genome is pervasively and extensively transcribed and encodes several thousand non-protein-coding RNAs (ncRNAs), and (iii) sequencing all these RNA molecules and making in silico alignments with the DNA genomic sequence indicates that up to 90 % of the euchromatic genome of the mouse is transcribed, sometimes from both DNA strands, and in both directions (many sense–antisense pairs).

Nowadays, the mammalian genome can no longer be regarded as a mere repository of the basic information necessary for the synthesis of thousands of proteins, but rather as a sophisticated factory releasing a great variety of coding and non-coding RNAs (ncRNAs) of various sizes and functions. In spite of enormous progress in the sequencing technology of nucleic acids, the inventory of these molecules is far from being completed and their annotation may still require several years. It has been established, for example, that many primary RNA transcripts are processed into smaller sized molecules, while others are alternatively spliced, thus tremendously increasing the complexity and diversity of the transcriptome. For this reason, scientists sometimes refer to this new category of non-coding RNAs as "*the dark matter of the transcriptome*". We will summarize the situation as it stands presently based on recent reviews on the subject, but it is clear that this chapter, more than any others in this book, will require regular updating. Undertaking the exhaustive inventory of the ncRNAs encoded in the mouse genome and performing their annotation is nothing less than embarking on the exploration of "*a new continent in the RNA world*".

5.4.1 ncRNAs Involved in Protein Synthesis

In addition to the messenger RNAs (mRNAs), which are protein-coding and are considered as the "noble" RNAs since they represent the message transcribed from the DNA, four types of ncRNAs have been described as essential components in the successive steps of protein synthesis and processing: transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), short non-coding RNAs (snRNAs, sometimes referred to as U-RNAs) and small nucleolar RNAs (snoRNAs).

5.4.1.1 Transfer RNAs

Transfer RNAs are a relatively homogeneous family of "adaptor" molecules whose function is to mediate recognition of a specific codon in the processed mRNA and to provide the corresponding amino acid during the process of protein synthesis (elongation step). These RNAs are part of a larger transcript, the pre-tRNAs, in the nucleus, which are subsequently split into smaller molecules (average 80 nucleotides), with a typical 3D cloverleaf structure that is well adapted to the function since, as we said, the tRNA binds to the mRNA codon (specific), to the new incoming amino acid (specific), and to the last amino acid incorporated into the growing polypeptidic chain. There are around 500 tRNA-encoding genes in the mouse genome and about the same number of pseudogenes. The tRNA-encoding genes are dispersed over the whole genetic map, including both the X and the Y chromosomes. Their computerized prediction is difficult due to their short size and, mostly, to the existence in the mouse genome of a very high number of short interspersed sequences (SINEs—see above) that originally derived from tRNA genes but are now inactive. This is typically a source of confusion and, for this reason, the experimental verification of the status of a potentially novel tRNA gene must be part of the annotation process (Coughlin et al. 2009).

5.4.1.2 Ribosomal RNAs

In contrast with the tRNAs, ribosomal RNAs are a relatively heterogeneous family of molecules with a size between 150 and ~5,000 nucleotides. The family comprises four types of RNAs (28S, 5.8S, 5S, and 18S). The 28S RNA (5,070 nt) and the 5.8S RNA (156 nt) bind to each other and are associated with the 5S RNA (121 nt) and with at least 45 proteins, to make the ribosomal large unit (60S). The 18S rRNA (comprising 1,869 nt) is associated with around 33 proteins to make the ribosomal small unit (40S). The two ribosomal subunits, the small and the large, are tightly associated to make the cytoplasmic ribosomes. The biosynthesis of mature ribosomes is complex and involves numerous processing events with the participation of other ncRNAs. When mature, the ribosomes serve as workbenches for protein synthesis. The mRNA is held sandwiched between the two subunits of

the rRNAs while being "scanned" and then transcoded into proteins. rRNAs are rapidly degraded in the cytoplasm once they have been used for protein synthesis. The genes encoding ribosomal RNAs are very numerous and spread over the whole genome (Henderson et al. 1974). They are organized in repeated units that, in the mouse, are 44 kb long. Each repeat contains three of the genes encoding rRNA, namely 18S, 5.8S, and 28S, and constitutes a transcription unit producing polycistronic RNA that is cleaved apart afterwards. These units are tandemly repeated and constitute the so-called nucleolar organizers (or NORs). These are distributed over several chromosomes (Chrs 4, 12, 15, 16, 18 and 19) in the case of *Mus m. domesticus*, but on all 40 chromosomes except the Y in *Mus caroli* (Rowe et al. 1996; Cazaux et al. 2011). At the end of mitosis (telophase) when rDNA transcription by RNA Polymerase I resumes, the NORs gather in the nucleolus (a nuclear organelle where rRNAs are produced and assembled with ribosomal proteins to form functional ribosomes). Genes that encode rRNA are expressed in virtually all types of cells and in all species, including prokaryotes. For this reason, many rRNAs have been sequenced and their sequences are now used as tools for systematics (ribotyping).

5.4.1.3 Small Nuclear RNAs

Small nuclear RNA molecules are found in the nucleus of eukaryotic cells. As is the case for many other small-sized RNAs, they are transcribed as larger molecules that are cleaved afterwards. They have an average length of approximately 150 nucleotides and are generally classified into five categories: U1, U2, U4, U5, and U6. Each of these snRNAs is associated with a large set of specific proteins (over 150), and the complexes they form with these proteins are referred to as small nuclear ribonucleoproteins (snRNPs or "snurps"). The snurps are essential in the splicing process. The splicing of mRNAs is a very complex and extremely precise process and this is probably why the spliceosome requires so many components to make its functioning totally error-proof. Each of the five categories of snRNAs has specific binding sequences and a specific function on the pre-mRNA substrate.

5.4.1.4 Small Nucleolar RNAs

The small nucleolar RNAs are small molecules measuring 60–300 nt. They are involved in the processing of rRNAs and are essential for ribosome maturation. They can also regulate the splicing of some mRNAs by modifying small nuclear RNAs (snRNAs) that are the major RNA component of the spliceosome, as we mentioned. snoRNAs probably have many other functions that have not yet been described, and the inventory of this family of molecules is difficult because their computerized prediction and classification is unreliable, yielding many orphan snoRNAs. snoRNAs encoding genes have been identified at several loci in the

mouse genome (2, 7, 8, 9, 12, 17, and X). The range of functions of these RNAs is likely to expand with the discovery of new molecules (Gardner et al. 2010).

Some genetic diseases affecting humans (for example spinal muscular atrophy and congenital dyskeratosis) have been correlated to abnormal functioning of the snurps. Prader–Willi syndrome (and the reciprocal Angelman syndrome—see Chap. 6 for details) is caused by the abnormal imprinting of a cluster of snoRNAs encoding genes located in the q11-13 region of human chromosome 15 that are involved in the synthesis of the serotonin-2C receptor mRNA. snRNAs also play an important role in maintaining the size of the telomeres (see Chap. 3).

5.4.2 The ncRNAs Functioning as Post-transcriptional Regulators

5.4.2.1 MicroRNAs

MicroRNAs (miRNAs) are small, single-stranded RNAs, measuring 21–24 nt (average 22 nt), whose function is to negatively regulate specific genes by mRNA degradation or translational repression. Around 60 % of these miRNAs are encoded in the intergenic regions and in antisense orientation to certain genes, and 40 % are encoded in the intronic regions of genes encoding proteins. These RNAs (along with the small interfering RNAs, described later) are the most well-known family of non-protein-coding RNAs.

The DNA encoding miRNAs is transcribed into precursors called pri-miRNAs. Each of these pri-miRNAs folds to form a double-stranded structure by base-pairing with itself. This structure looks like a hairpin with a few loops of stranded RNA. The pri-miRNA is then cleaved into a precursor known as a pre-miRNA, which is transported into the cytoplasm. Finally, the pre-miRNA is incorporated into a molecular complex of proteins of the argonaute family called the *miRNA-induced silencing complex* or *miRISC*. The processing of mature miRNAs requires the participation of an endoribonuclease known as *Dicer* that cleaves the pre-miRNA into the mature miRNA. miRISC modulates the activity of the targeted mRNA by identifying a 2–7-bp complementary sequence, known as the "*seed region*", which is generally located at the 3'-UTR. Both the processing and the loading of miRNAs into the RISC complex and the function of this machinery are precisely regulated (Ebert and Sharp 2012).

The fact that these miRNAs exist in several species including invertebrates and plants, and the way they are transcribed and processed from highly preserved sequences, with highly sophisticated mechanisms, indicates that they probably represent an ancestral mechanism of gene regulation (Lewis and Steel 2010). Because they also have a wide range of spatial and temporal expression patterns, they probably play important roles at different steps of embryonic development and in some pathological conditions. Indeed, it is expected that about 60 % of mammalian protein-coding genes are more or less regulated by miRNAs.

miRNAs are numerous and distributed throughout the genomes of both animals and plants. In the mouse, as in humans, their number has been estimated in the range of 1,000. miRNAs are involved in many regulation processes, including cell proliferation, differentiation, apoptosis, and development. They function via base-pairing with complementary sequences of mRNA molecules (seed region), leading either to translational repression or to silencing via target degradation.

miRNA nomenclature consists of the generic or root symbol *Mir*, followed by the numbering in the miRBase database (www.mirbase.org), a database that tracks microRNAs reported for all species. Mouse *Mir143* (microRNA 143), for example, is represented as mmu-mir-143 in miRBase, with the mmu signifying *Mus musculus* (Fig. 5.12).

Demonstration of the involvement of miRNAs in a given developmental or pathological process is not easy. In the mouse, this can be achieved, for example, by performing the complete elimination of all miRNAs in a certain tissue or cell type and then observing the phenotypic effects. Since the *Dicer* protein is essential for the processing of miRNAs, as discussed above, mice with a conditional knockout allele of *Dicer* targeted in Purkinje cells (see Chap. 8—targeted knockout) no longer had any miRNAs in these cells, and were found to develop ataxia with Purkinje cell degeneration. This indicates that at least some miRNAs are indispensable for the differentiation of these highly specific cells (Schaefer et al. 2007). Another more specific strategy would be to establish an indisputable causal and direct relationship between a point mutation in the sequence of a given miRNA and a particular phenotype. Examples of this type are now accumulating,

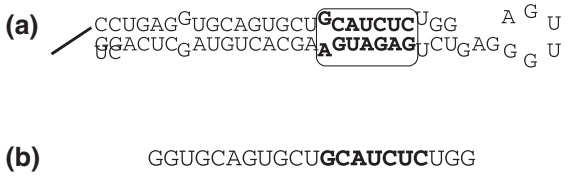


Fig. 5.12 *The microRNAs.* MicroRNAs (miRNAs) are short, noncoding, single-stranded RNAs. These miRNAs are nested within longer non-coding RNA molecules, which are processed in several successive steps with a double-stranded pre-miRNA (**a**), and finally a functional single-stranded RNA molecule measuring 20–22 bp (**b**). These miRNAs finely regulate the expression levels of several genes by binding to the 3'-untranslated regions of the corresponding mRNAs. The seed sequence of miR-143 (represented in *bold*) matches perfectly with the 3'-UTR of the mRNA transcribed from the (cytosine-5)-methyltransferase 3A (*DNMT3A*) gene. mir-143 is known to be involved in cardiac morphogenesis, it has also been implicated in human colon cancer development, and its expression is down-regulated during mouse odontoblast differentiation. mir–143 is encoded in mouse Chr 18 and is transcribed from the same DNA as another miRNA (mir-145). miRNAs are highly conserved in vertebrates, and this is suggestive of an important function. It is expected that about 60 % of mammalian protein-coding genes are more or less regulated by miRNAs

and one of the first and most well-documented cases is the semidominant mutation *Diminuendo* (symbol *Mir96*^{Dmdo}-Chr 6) (Lewis et al. 2009; Lewis and Steel 2010). This mutation was observed in the progeny of a male treated with the chemical mutagen ENU (see Chap. 7) and was presumably induced by this substance. The phenotype is characterized by progressive deafness, a condition that is quite common in humans. After positional cloning and careful sequencing of several candidate DNA segments in the 4.96-Mb critical interval where *Diminuendo* was mapped, the researchers finally found an A →T transversion in the "seed" region of the miRNA *Mir96*. This mutation, which was unique to *Diminuendo* and absent in all other mice as well as in a large series of vertebrates, was confirmed as the causative agent of the deafness and was associated with the down-regulation of several (at least five) proteins, each of them being involved in the function of the hair cells of the inner ear. These five proteins, which are downstream in the cascade of regulation initiated by *Mir96*, are all important for the differentiation and function of the hair cells and were all found to result in deafness when individually knocked out.

The discovery of the molecular origin of the *Diminuendo* mutation is an example of the role that the myriad of miRNAs may play in the fine regulation of gene (mRNA) expression in several developmental or pathological processes in vertebrates. The discovery of a point mutation in the seed region of *Mir96* proved that cell differentiation and organogenesis involve a network of functionally linked proteins as well as one or several miRNA(s).

Identification of the miRNA targets would certainly represent an enormous step forward in developmental genetics, and this is therefore a focus in many laboratories worldwide. Progress, however, is hampered by the fact that miRNAs are very small molecules and their sequences are not often totally complementary to their targets. In addition to this difficulty, many scientists also believe that many mRNAs, if not all, offer several targets to several miRNAs in their 3'-UTRs, thus adding even more complexity to the picture.

MicroRNAs definitely have a promising future in medicine because they are simple molecules but have, at the same time, the power of interfering with gene regulation. In humans they are intensively studied because their expression levels have been found increased in certain forms of cancers (for example, lymphomas or chronic lymphocytic leukemias), in diseases like cardiomyopathies, and in some infectious diseases or autoimmune diseases. These increases in specific miRNAs can then be used as information for the diagnosis or prognosis of the disease, or as potential treatments. For example, aortic banding in mice induces cardiac hypertrophy and concomitant up-regulation of many (over 100) miRNAs including *Mir21*. When *Mir21* was knocked down using an antisense approach, cardiomyocyte hypertrophy was reduced, suggesting that this particular miRNA plays a key role in the mechanism of cardiac hypertrophy. This obviously opens perspectives for the development of novel therapies.

Scientists believe that there are different grades in the process of mRNA regulation by miRNAs. Some miRNAs regulate specific individual targets, but it

seems that key miRNAs (so-called "super-miRNAs") can regulate the expression levels of hundreds of genes simultaneously and cooperatively. These super-miRNAs are of course actively searched. It has been suggested that miRNAs exert both absolute and fine-tuned control of gene expression, adjusting levels of transcripts to give either complete repression or simply decreased expression. Such "fine-tuning" miRNAs will be much harder to identify than those resulting in the complete "switching off" of a gene, since loss of function of any of these miRNAs would presumably have subtle effects, which would be difficult to characterize and study.

The discovery over the past ten years of these post-transcriptional regulators has opened up a "*new continent of the RNA world*". We just gave a rapid overview of this continent using the miRNAs as examples, but many other RNA or RNA-like molecules are just as interesting. We will now consider the case of siRNAs, another type of ncRNA with post-transcriptional regulatory functions.

5.4.2.2 Small Interfering RNA

Small interfering RNA, short interfering RNA, or silencing RNAs (all abbreviated siRNAs) are short double-stranded RNA molecules (20–25 bp) with a 2-bp 3' overhang and phosphate groups on the 5' end of each strand. These RNAs interfere with (i.e., reduce or suppress) the expression of specific genes with complementary nucleotide sequence, and in so doing they obviously have similarities in their mode of action with the miRNAs discussed above.

The existence of these siRNAs and their remarkable properties were discovered by chance while plant geneticists were performing transgenic experiments with the aim of darkening the color of petunia flowers. The transgene they were using was that for chalcone synthase, a key enzyme of the flavonoid/isoflavonoid biosynthesis pathway. The scientists expected that by increasing the enzyme level with several extra transgenic copies of the gene, this may influence the pigmentation of the flower (Napoli et al. 1990). In fact, and to their surprise, instead of obtaining the dark purple flowers they expected they got light-colored flowers and sometimes flowers with white (unpigmented) patches, indicating that the chalcone-encoding transgene actually had adverse effects on the pigmentation process. Other similar experiments revealed that the observed phenotypes were not exceptional but, on the contrary, the consequence of an increased rate of mRNA degradation leading to specific gene suppression or, more precisely, down-regulation. This effect was designated RNA interference or RNAi.

In 1998, Fire and colleagues (1998), performing a similar experiment with the worm *Caenorhabditis elegans,* concluded that neither the complete mRNA nor a variety of antisense RNAs had an effect on protein production in experimentally injected worms. However, they found that double-stranded RNAs corresponding to a myofilament protein successfully silenced the targeted genes,

once injected under the same conditions. They also demonstrated that only a few molecules of injected double-stranded RNA were required to induce gene silencing, thus arguing against stoichiometric interference with endogenous mRNA and suggesting that there could be a catalytic or amplification component in the interference process. This finding had a great impact in biology and medicine when it was demonstrated that RNAi mechanisms are universal and active in humans as well as in several model organisms including rats and mice, offering new tools for gene annotation as well as opening the way to the development of novel therapeutic strategies for the treatment of genetic diseases, including cancers.[8]

Unlike in many model species, RNA interference cannot be triggered in mammalian cells by injecting long double-stranded RNAs, because the cells recognize these RNAs as viruses and immediately develop a deleterious interferon response with consequences for cell survival. Short molecules do not trigger this reaction when injected into the cells.

siRNAs can also be synthesized as single-stranded molecules in the laboratory and then introduced into cells either by direct injection or by transfection. Direct chemical synthesis has the great advantage of allowing slight variations in the sequence, and as a result increasing the efficiency of the siRNAs. Not all native siRNAs are equally active, and the possibility of synthesizing novel molecules appears to be a promising strategy (Ramachandran and Ignacimuthu 2013). The mechanisms by which miRNAs and siRNAs work are similar. However, while miRNAs cause translational repression or destabilization, the siRNAs cleave their target RNAs at a particular site.

The use of RNA interference is an interesting and efficient way of altering the gene function and accordingly of performing gene annotation. However, in most instances and unlike other strategies described in Chap. 8, RNA interference induces down-regulation of gene expression (knockdown) and not knockout proper. In addition, some of these knockdowns are not specific.

5.4.2.3 Piwi-Interacting RNAs

Piwi-interacting RNAs (*piRNAs*) are short ncRNAs (26–31 nt long), which are expressed mainly, not to say specifically, in spermatogenic cells of mammals. Their function is not yet fully understood, but it is known that they form complexes with the regulatory piwi (or miwi) proteins. These piRNA complexes are thought to play a role in transposon silencing in male germ line cells, limiting the expansion of these repeated sequences. They presumably have other functions that have not yet been characterized.

[8] A. Fire and C. Mello were awarded the Nobel Prize in Physiology or Medicine in 2006 for their discovery of "RNA interference—gene silencing by double-stranded RNA".

5.4.2.4 Long Non-coding RNAs

Long non-coding RNAs (lncRNAs) have an average size larger than 200 nt and in many cases, in the range of 2 kb or more. This relatively great size distinguishes them from all other ncRNAs, but being similar in size to the mRNAs can hamper their isolation and characterization. Computer algorithms assessing the coding potential of the two molecules (lncRNAs and mRNAs) have been used to discriminate between these molecules when necessary, but this criterion has finally proven unreliable because some (not all) lncRNAs do have a coding frame or, more precisely, a nucleotide sequence resembling a coding frame with start and termination codons. So far, the analysis of the sequences of lncRNAs does not allow sorting them in discrete families with specific functions. In addition, the sequences of these RNAs are only poorly conserved across species, even among closely related mammals. Indeed, this family of ncRNAs is heterogeneous to the point where its very existence has long been debated. Since lncRNAs are four times more numerous than mRNAs, one can understand why they have been designated the "dark matter" of the transcriptome.

Aside from this rather confusing situation, some data have recently emerged that make the situation a little more coherent. First, sequence alignments reveal that lncRNAs are transcribed from both strands and in both directions overlapping introns, sometimes exons, and intergenic regions: this is never the case with mRNAs. Also, unlike mRNAs, many of these molecules stay in the nucleus, suggesting that they have a function at or close to this location. Finally, and as we will discuss further, the density of lncRNAs seems to be locally associated with some pathologies, suggesting that they may be involved more or less directly in these processes.

Most of the knowledge we have of the lncRNAs results from the studies of five important lncRNAs that have been studied in the mouse and whose functions have now been relatively well characterized: these are the *Kcnq1* overlapping transcript 1 (*Kcnq1ot1*-Chr 7), the antisense IGF2R-RNA (*Airn*-Chr 17), the HOX transcript antisense RNA (*Hotair*-Chr 15), the X-specific transcripts (*Xist*-Chr X), and the X (inactive)-specific transcript, antisense (*Tsix*-Chr X). The function and mode of action of the lncRNAs involved in the X-chromosome inactivation process will be analyzed in Chap. 11. *Xist* is one of the first genes, expressed after fertilization, leading to silencing of all the genes on the targeted chromosome as a consequence of histone H3 modifications. The targeting of XIST RNA to only one of the chromosomes is controlled by another lncRNA: TSIX, which is the antisense repressor of *Xist* on the active X chromosome.

Antisense repression is also the mode of action of the gene *Kcnq1,* whose expression is silenced by the paternally expressed antisense non-coding RNA KCNQ1OT1.

LncRNAs have extremely variable stability and expression levels. Some have a half-life of only one hour (for example, KCNQ1OT1), while others are much more stable. Some are highly expressed, while others are barely detectable.

Indeed, from the many reviews that have been published, one can conclude that "*we have barely begun to scratch the surface of the lncRNA world*" (Kung et al. 2013).

5.5 Ultraconserved Elements (UCE) and Long Conserved Non-coding Sequences

When the mouse genomic sequence is aligned to the genomic sequence of other vertebrate species, we observe that quite a large number of elements measuring ≥200 bp are conserved, and sometimes highly conserved. These sequence elements are commonly designated *ultraconserved elements* (UCEs). UCEs were first described in the human, rat, and mouse genomes by Bejerano and coworkers (2004), but were also discovered in many other more distantly related species (chicken, for example). For the UCEs encoding proteins or functional RNAs, geneticists have an explanation: they consider that these resemblances are a consequence of strong selection pressures acting during evolution and that we mentioned earlier as "genome-shaping forces". However, the situation is much less clear for the non-protein-coding UCEs, and in this case explanations are lacking.

After alignment of the mouse and human genomes, scientists at the RIKEN Institute identified over 600 such conserved non-coding DNA sequences with nearly 95 % identity and a size greater than 500 bp, most of them independent of the previously reported UCEs (Sakuraba et al. 2008). These sequences, which they provisionally designated *long conserved non-coding sequences* (LCNS), were also found scattered throughout the genome of the rat as well as other vertebrate species (chick, frog, fish) but were not found in non-vertebrate species. Given that the probability of finding sequence similarities of that kind, just by chance, is extremely low, two hypotheses were proposed by the researchers to account for their observations: the first hypothesis was to consider that these LCNS either have an important although unknown function associated with their structure (they could have regulatory or structural elements important for the chromosome structural organization, for example), or that they are transcribed into functional ncRNAs whose function is not yet established (perhaps a type of lncRNA); in both cases, this would explain why the sequences in question were selectively constrained. The second hypothesis is that the LCNS/UCEs have remained intact for so many years of evolution, simply because they are mutational cold spots (Katzman et al. 2007). To challenge these hypotheses, the scientists had the clever idea of performing ENU mutagenesis and measuring, afterwards, the frequencies of induced mutations in the LCNS and comparing it with other genomic regions. They did not find any significant differences in the mutation rates after screening 40.7 Mb of conserved sequences (~35 mutations) and concluded that the LCNS were not mutational cold spots. To date, we do not have any satisfactory explanation to account for the presence of so many of these LCNS/UCEs. The scientists of the ENCODE project consider them to be associated with gene regulation (ENCODE Project 2012) and their role is probably essential if we consider their near-universal conservation across extremely divergent species. On the other hand, it has also been reported that deletions of these UCEs in mice had virtually no effect on the viability or fertility of the animals (Ahituv et al. 2007). This indicates

that extreme sequence constraints do not necessarily correspond to crucial functions. For mouse geneticists, this also indicates that another type of sequence element must now be added to the "*dark matter of the transcriptome*".

5.6 Mitochondrial DNA

Mitochondria have a genome of their own that is represented by a small, circular, double-stranded DNA molecule known as mtDNA, sometimes mDNA. In the mouse (as in humans) there are between two and ten such mtDNA molecules per mitochondrion, and the number of mitochondria per cell is extremely variable and depends on the cell type. The oocyte, for example, contains up to 10^6 mtDNA copies while a mature sperm cell contains less than 100.

The mtDNA comprises 37 genes encoding 13 mitochondrial enzymes involved in respiration and oxidative phosphorylation, two ribosomal RNAs (12S and 16S) and a full set of 22 tRNAs that are essential for the synthesis of these enzymes. However, this small set of proteins represents only a sampling of the ~1,500 mitochondrial proteins, the rest of them being encoded in the nuclear genome. In contrast to the mammalian nuclear DNA, mtDNA is a naked DNA molecule (i.e., histone-free) with no introns and no sequence repeats (Bayona-Bafaluy et al. 2003). In addition, its two strands are quite different from those in the nuclear DNA, the heavy strand being very heavy while the light one is much lighter. All these unique characteristics of the mtDNA molecule are generally correlated with its presumptive evolutionary origin, which states that the mitochondria are remnants of bacteria that have been incorporated into the primitive eukaryotic cells and retained as symbiotic organisms due to their selective advantages for cellular metabolism. This interesting hypothesis, which is also proposed for chloroplasts in plants, is not formally confirmed but it seems more than likely and fits perfectly with the molecular data accumulated recently, in particular some fundamental changes in codon usage[9] (Yu et al. 2009).

The consensus sequence of the mouse mtDNA has been established and found to consist of around 16,300 bp, with point variations (a few SNPs and gaps or indels) among the most common laboratory inbred strains and the most commonly used mouse species (Goios et al. 2007, 2008). These sequence polymorphisms have been cleverly exploited to establish or to confirm the phylogenetic relationships relationships between the different species of the genus *Mus* and related genus (see Chap. 1) and the historical phylogeny among the laboratory strains (see Chap. 9). This has allowed, in particular, the confirmation that a great majority

[9] There are a few differences between the vertebrate mtDNA code and the "universal" code. In the mtDNA, UGA codes for Trp rather than being a stop codon. In the same mtDNA there are two Met codons (AUA and AUG) rather than only one. Finally, both AGA and AGG are read as stop codons.

of the most frequently used inbred strains were all derived from the same female ancestor, as initially established by Yonekawa et al. (1982), and to confirm that most laboratory strains can be sorted into three groups with independent ancestral/geographical origins: the Sino-Japanese mice, the Swiss mice and the "Abbie Lathrop's" mice in the United States.

The mtDNA replicates at a much higher rate than the nuclear DNA and does not possess repair mechanisms as efficient as those of the latter. For this reason, and probably also because the mtDNA is not protected from the mutagenic action of its environment by a variety of histone proteins, as is the case for mammalian DNA, it is more "mutable" and appears to be about 10–20 times more affected by mutations generating a sequence polymorphism than the nuclear DNA of the same species. Considering the great differences between male and female gametes in terms of mitochondria numbers (up to 1/1,000), it is no surprise to learn that the mtDNA is transmitted by the mother to her offspring rather than by the father. Although sperm cells do have some mtDNA molecules, the mtDNA appears to be lost very early during egg development, and in virtually all species studied so far the only mtDNA molecules found in embryos are of maternal origin.

When a mutation occurs in a mtDNA molecule of an oocyte (or of a precursor cell), it is generally counter-selected and rapidly eliminated unless it confers a selective advantage to the mtDNA, for example by increasing its replication rate (Sharpley et al. 2012). In the latter case, the mutant molecules progressively overgrow the population of normal mtDNAs and the oocyte (or cell) becomes heteroplasmic with two (or more) types of mtDNA. Finally, due to some sort of sampling effect, sometimes referred to as a genetic bottleneck, the mutant form of the mtDNA may completely replace the pre-existing form and become the standard. This explains why mtDNA is an attractive molecule to geneticists studying evolution. It is also interesting to note that mtDNA evolution is completely independent of nuclear DNA evolution, and accordingly represents another valuable tool for establishing the systematics of a species. For this reason, it has been extensively used in many domestic species, including the mouse, and still is.

In the human species, mutations in the mtDNA have been associated with more or less severe pathologies. Leber hereditary optic neuropathy (LHON), for example, was the first reported and is one of the most prevalent, with an estimated frequency of 15 in 100,000 births. This syndrome is the consequence of mutations (several have been described) occurring in the genes encoding the oxidative phosphorylation complex I. Many other mtDNA defects have been reported in the human species, including a syndrome of maternally inherited diabetes and deafness (MIDD), Leigh syndrome, a syndrome associating neuropathy, ataxia, retinitis pigmentosa, and ptosis (NARP), myoneurogenic gastrointestinal encephalopathy (MNGIE), and many other neuromuscular diseases. All these pathologies are maternally transmitted and exhibit variations in severity presumably associated with the degree of heteroplasmy. Surprisingly, no such pathologies clearly attributable to an mtDNA defect have ever been reported in the mouse, although mtDNA mutations have been reported in cell lines transplanted in vitro.

Because spontaneous mtDNA defects have never been reported in the mouse, animal models of human pathologies have been created by introducing defective human mtDNAs into mouse oocytes.[10] In particular, a murine model of LHON syndrome has been produced by using this strategy. These mice exhibited reduction in retinal function, indicating that the physiopathology of the syndrome may result from some oxidative stress (Lin et al. 2012).

Those readers of this chapter who might be interested in the biology and pathology of mtDNA, in both human and mouse, should refer to the important contribution of D.C. Wallace (University of Pennsylvania), who wrote several reviews on the subject (Wallace 2009).

5.7 Conclusions

At the beginning of this chapter we stated that we considered the decision taken several years ago to completely and systematically sequence the mouse genome to be a wise one. If we consider the huge amount of information gathered, directly or indirectly, from this sequencing and the data we can expect to collect in the near future, our initial feeling is strengthened; indeed, the sequencing of the mouse genome has had an enormous impact in many areas of genetics and biology.

The knowledge of this sequence has allowed the development of better tools (for example, SNPs) and allows better experiments to be designed. Nowadays one can design an experiment of homologous recombination (targeted mutagenesis) with precision at the base-pair level.

Aside from these technical advances, in silico comparisons of the mouse sequence with other mammalian (or vertebrate) sequences has allowed the discovery of similarities or differences that have proved a rich source of information for a better understanding of evolution. Even within individuals of the same species, the analysis of copy number variations, for example, has revealed intriguing differences whose significance and phenotypic expression is not yet completely clear, even if we suspect that they probably play an important role in quantitative genetics.

The information gathered concerning the structure of the mouse genome and its variations across the different inbred strains and different subspecies of the *Mus* genus will certainly reveal important clues for understanding the genetic determinism of complex traits, especially when complemented by the constantly increasing amounts of phenotyping data. The mouse is unique in the sense that one can cross animals of different subspecies, breed very large progenies, extensively phenotype all the animals, and sequence the individual genomes when desired.

[10] Two inbred strains of mice with the same genomic (nuclear) DNA but different mtDNAs are said to be conplasmic. The production of such strains can be achieved by normal sexual reproduction or by direct cytoplasmic transfer (See Chap. 9).

The sequencing of the genome has also revealed its great plasticity. We now know that LINEs and ERVs play an important role in gene regulation, and even as a source of diversity, a point that was totally unexpected.

Finally, a true revolution in our understanding of the transcriptome occurred during the last ten years. The number of protein-encoding genes has been revised downward while the number of RNA-encoding genes is constantly being revised upward. Over the last ten years we have started to realize that a myriad of ncRNAs (long and short) are transcribed from the genome, exhibiting great although still incompletely explored functional diversity. From whole-genome analyses using microarrays and high-throughput transcript sequencing, we estimate that more than 85 % of the nucleotides in the euchromatic genome are represented in primary transcripts. Indeed, the proportion of supposedly "junk" DNA shrinks more every day. We have learnt that the genome is pervasively and bidirectionally transcribed, increasing tremendously the amount of information that can be stored. The discovery of the role of miRNAs and siRNAs in the fine regulation of gene activity is another revolution that may have major consequences for the diagnostic and treatment of some diseases. The long coding RNAs probably play a major role in gene regulation and imprinting … but we have information about only a handful of these molecules although we know that there are many.

The role and importance of the ultraconserved elements and long conserved non-coding sequences remains a mystery. If they are ultraconserved this would mean that they are under selection pressure. But, alternatively, we know that they can be experimentally deleted with apparently no consequences. No doubt all these observations will fuel much research in the years to come and it won't be surprising that, at this point, even the concept of gene may be reconsidered[11].

Acknowledgements The authors thank Doctor Benoît Robert, Institut Pasteur, for his contribution to Sect. 5.3.3 of this chapter.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM (2007) Deletion of ultraconserved elements yields viable mice. PLoS Biol 5:e234

Arnold CN, Xia Y, Lin P, Ross C, Schwander M, Smart NG, Müller U, Beutler B (2011) Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. Genetics 187:633–641

Balakirev ES, Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet 37:123–151

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. Nature 465:53–59

[11] Most of the data provided in this chapter concerning the Mouse Genome are from the Ensembl website: http://www.ensembl.org/Mus_musculus/Info/Annotation#assembly and http://www.ncbi.nlm.nih.gov/projects/mapview/stats/BuildStats.cgi?taxid=10090&build=38&ver=1.

Bayona-Bafaluy MP, Acín-Pérez R, Mullikin JC, Park JS, Moreno-Loshuertos R, Hu P, Pérez-Martos A, Fernández-Silva P, Bai Y, Enríquez JA (2003) Revisiting the mouse mitochondrial DNA sequence. Nucleic Acids Res 31:5349–5355

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325

Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P (2010) Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res 38:3909–3922

Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L, Gurrieri F, Everman DB, Schwartz CE, Birk OS, Bejerano G, Lomvardas S, Ahituv N (2012) Coding exons function as tissue-specific enhancers of nearby genes. Genome Res 22:1059–1068

Blanco E, Guigo R (2005) Predictive methods using DNA sequences—analysis at the nucleotide level. In: Baxevanis AD, Francis Ouellette BF (eds) Bioinformatics: a practical guide to the analysis of genes and proteins, 3rd edn. Wiley, Hoboken

Buckler AJ, Chang DD, Graw SL, Brook JD, Haber DA, Sharp PA, Housman DE (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc Natl Acad Sci U S A. 88:4005–4009

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Cachon-Gonzalez MB, Fenner S, Coffin JM, Moran C, Best S, Stoye JP (1994) Structure and expression of the hairless gene of mice. Proc Natl Acad Sci U S A 91:7717–7721

Canales CP, Walz K (2011) Copy number variation and susceptibility to complex traits. EMBO Mol Med. 3:1–4

Carlson CM, Largaespada DA (2005) Insertional mutagenesis in mice: new perspectives and tools. Nat Rev Genet 6:568–580

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635

Cazaux B, Catalan J, Veyrunes F, Douzery EJ, Britton-Davidian J (2011) Are ribosomal DNA clusters rearrangement hotspots?: a case study in the genus Mus (Rodentia, Muridae). BMC Evol Biol 11:124. doi:10.1186/1471-2148-11-124

Cook EH, Scherer SW (2008) Copy-number variations associated with neuropsychiatric conditions. Nature 455:919–923

Copeland NG, Jenkins NA (2010) Harnessing transposons for cancer genes discovery. Nat Rev Cancer 10:696–706

Coughlin DJ, Babak T, Nihranz C, Hughes TR, Engelke DR (2009) Prediction and verification of mouse tRNA gene families. RNA Biol 6:195–202

Cutler G, Kassner PD (2008) Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. Cytogenet Genome Res 123:297–306

Demuth JP, Hahn MW (2009) The life and death of gene families. BioEssays 31:29–39

Ebert MS, Sharp PA (2012) Roles for microRNAs in conferring robustness to biological processes. Cell 149:515–524

Egan CM, Sridhar S, Wigler M, Hall I (2007) Recurrent DNA copy number variation in the laboratory mouse. Nat Genet 39:1384–1389

Ehrnhoefer DE, Butland SL, Pouladi MA, Hayden MR (2009) Mouse models of Huntington disease: variations on a theme. Dis Model Mech 2:123–129

ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 4:e1001046

ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811

Gardner P, Bateman A, Poole AM (2010) SnoPatrol: how many snoRNA genes are there? J Biol 9:1–4

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521

Goios A, Pereira L, Bogue M, Macaulay V, Amorim A (2007) mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res 17:293–298

Goios A, Gusmão L, Rocha AM, Fonseca A, Pereira L, Bogue M, Amorim A (2008) Identification of mouse inbred strains through mitochondrial DNA single-nucleotide extension. Electrophoresis 29:4795–4802

Goldberg ML. 1979. PhD Diss. Stanford University, Stanford, CA

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440

Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, Hayashizaki Y, Carninci P (2006) The complexity of the mammalian transcriptome. J Physiol 575:321–332

Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 13:469–483

Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R (2009) Identifying protein-coding genes in genomic sequences. Genome Biol 10:201 Epub

Hatzis P, van der Flier LG, van Driel MA, Guryev V, Nielsen F, Denissov S, Nijman IJ, Koster J, Santo EE, Welboren W, Versteeg R, Cuppen E, van de Wetering M, Clevers H, Stunnenberg HG (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol Cell Biol 28:2732–2744

Hayashizaki Y, Carninci P (2006) Genome Network and FANTOM3: Assessing the Complexity of the Transcriptome. PLoS Genet 2(4):e63

Henderson AS, Eicher EM, Yu MT, Atwood KC (1974) The chromosomal location of ribosomal DNA in the mouse. Chromosoma 49:155–160

Hill RE (2007) How to make a zone of polarizing activity: insights into limb development via the abnormality preaxial polydactyly. Dev Growth Differ 49:439–448

Hoskins AA, Moore MJ (2012) The spliceosome: a flexible, reversible macromolecular machine. Trends Biochem Sci 37:179–188

Howell VM (2012) Sleeping beauty—a mouse model for all cancers? Cancer Lett 317:1–8

International Human Genome Sequencing Consortium, Lander E, Linton L, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409:890–921

Izsvák Z, Ivics Z (2005) Sleeping Beauty hits them all: transposon-mediated saturation mutagenesis in the mouse germline. Nat Methods 2:735–736

Julier C, de Gouyon B, Georges M, Guénet JL, Nakamura Y, Avner P, Lathrop GM (1990) Minisatellite linkage maps in the mouse by cross-hybridization with human probes containing tandem repeats. Proc Natl Acad Sci U S A. 87:4585–4589

Kapranov P, St Laurent G (2012) Dark matter RNA: existence, function, and controversy. Front Genet. 3:60

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564–1566

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D (2007) Human genome ultraconserved elements are ultraselected. Science 317:915

Kozak M (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol 196:947–950

Kung JT, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193:651–669

Kuznetsova IS, Prusov AN, Enukashvily NI, Podgornaya OI (2005) New types of mouse centromeric satellite DNAs. Chromosome Res 13:9–25

Lagha M, Bothma JP, Levine M (2012) Mechanisms of transcriptional precision in animal development. Trends Genet 28:409–416

Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature 494:497–501

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A. 99:7548–7553

Lewis MA, Steel KP (2010) microRNAS in mouse development and diseases. Semin Cell Develop Biol 21:774–780

Lewis MA, Quint E, Glazier AM, Fuchs H, De Angelis MH, Langford C, van Dongen S, Abreu-Goodger C, Piipari M, Redshaw N, Dalmay T, Moreno-Pelayo MA, Enright AJ, Steel KP (2009) An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat Genet 41:614–618

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei CL, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, Fullwood MJ, Cheung E, Liu E, Sung WK, Snyder M, Ruan Y (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148:84–98

Lin CS, Sharpley MS, Fan W, Waymire KG, Sadun AA, Carelli V, Ross-Cisneros FN, Baciu P, Sung E, McManus MJ, Pan BX, Gil DW, Macgregor GR, Wallace DC (2012) Mouse mtDNA mutant model of Leber hereditary optic neuropathy. Proc Natl Acad Sci U S A. 109:20065–20070

Mashimo T, Glaser P, Lucas M, Simon-Chazottes D, Ceccaldi PE, Montagutelli X, Desprès P, Guénet JL (2003) Structural and functional genomics and evolutionary relationships in the cluster of genes encoding murine 2′,5'-oligoadenylate synthetases. Genomics 82:537–552

Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15 Spec No 1:R17–29

Modrek B, Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34:177–180

Mouse ENCODE Consortium (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13:418

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Mülhardt C, Fischer M, Gass P, Simon-Chazottes D, Guénet JL, Kuhse J, Betz H, Becker CM (1994) The spastic mouse: aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. Neuron 13:1003–1015

Munroe R, Schimenti J (2009) Mutagenesis of mouse embryonic stem cells with ethylmethanesulfonate. Methods Mol Biol 530:131–138

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J et al (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 296:1661–1671

Napoli C, Lemieux C, Jorgensen R (1990) Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. Plant Cell 2:279–289

Nouvel P (1994) The mammalian genome shaping activity of reverse transcriptase. Genetica 93:191–201

Ohno S (1972) So much "junk" DNA in our genome. In: Smith HH (ed) Evolution of genetic systems. Gordon and Breach, New York, pp 366–370

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I et al (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573

Panthier JJ, Dreyfus M, Roux TL, Rougeon F (1984) Mouse kidney and submaxillary gland renin genes differ in their 5' putative regulatory sequences. Proc Natl Acad Sci U S A. 81:5489–5493

Perelygin AA, Zharkikh AA, Scherbik SV, Brinton MA (2006) The mammalian 2′-5' oligoadenylate synthetase gene family: evidence for concerted evolution of paralogous Oas1 genes in Rodentia and Artiodactyla. J Mol Evol 63:562–576

Perez CJ, Dumas A, Vallières L, Guénet JL, Benavides F (2013) Several classical mouse inbred strains, including DBA/2, NOD/Lt, FVB/N, and SJL/J, carry a putative loss-of-function allele of Gpr84. J Hered 104:565–571

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K (2007) Evidence of a large-scale functional organization of Mammalian chromosomes. PLoS Biol 5(5):e127

Ramachandran PV, Ignacimuthu S (2013) RNA interference-a silent but an efficient therapeutic tool. Appl Biochem Biotechnol 169:1774–1789

Roberts RL, Diaz-Gallo LM, Barclay ML, Gómez-García M, Cardeña C, Merriman TR, Gearry RB, Martin J (2012) Independent replication of an association of CNVR7113.6 with Crohn's disease in Caucasians. Inflamm Bowel Dis 18:305–311

Rowe LB, Janaswami PM, Barter ME, Birkenmeier EH (1996) Genetic mapping of 18S ribosomal RNA-related loci to mouse chromosomes 5, 6, 9, 12, 17, 18, 19, and X. Mamm Genome 12:886–889

Sakuraba Y, Kimura T, Masuya H, Noguchi H, Sezutsu H, Takahasi KR, Toyoda A, Fukumura R, Murata T, Sakaki Y, Yamamura M, Wakana S, Noda T, Shiroishi T, Gondo Y (2008) Identification and characterization of new long conserved noncoding sequences in vertebrates. Mamm Genome 19:703–712

Savarese F, Grosschedl R (2006) Blurring cis and trans in gene regulation. Cell 126:248–250

Saxena A, Carninci P (2011) Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. BioEssays 33:830–839

Schaefer A, O'Carroll D, Tan CL, Hillman D, Sugimori M, Llinas R et al (2007) Cerebellar neurodegeneration in the absence of microRNAs. J Exp Med 204:1553–1558

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung

W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M (2007) Strong association of de novo copy number mutations with autism. Science 316:445–449

Sharpley MS, Marciniak C, Eckel-Mahan K, McManus M, Crimi M, Waymire K, Lin CS, Masubuchi S, Friend N, Koike M, Chalkia D, MacGregor G, Sassone-Corsi P, Wallace DC (2012) Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. Cell 151:333–343

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. Nature 488:116–120

Silver LM (1995) Mouse genetics—concepts and applications. Oxford University Press, Oxford

Sookdeo A, Hepp CM, McClure MA, Boissinot S (2013) Revisiting the evolution of mouse LINE-1 in the genomic era. Mob DNA 4:3. doi:10.1186/1759-8753-4-3

Specht CG, Schoepfer R (2001) Deletion of the alpha-synuclein locus in a subpopulation of C57BL/6 J inbred mice. BMC Neurosci 2:11

Stoye JP, Fenner S, Greenoak GE, Moran C, Coffin JM (1988) Role of endogenous retroviruses as mutagens: the hairless mutation of mice. Cell 54:383–391

Valentijn LJ, Baas F, Wolterman RA, Hoogendijk JE, van den Bosch NH, Zorn I, Gabreëls-Festen AW, de Visser M, Bolhuis PA (1992a) Identical point mutations of PMP-22 in Trembler-J mouse and Charcot-Marie-Tooth disease type 1A. Nat Genet 4:288–291

Valentijn LJ, Bolhuis PA, Zorn I, Hoogendijk JE, van den Bosch N, Hensels GW, Stanton VP Jr, Housman DE, Fischbeck KH, Ross DA et al (1992b) The peripheral myelin gene PMP-22/GAS-3 is duplicated in Charcot-Marie-Tooth disease type 1A. Nat Genet 1:166–170

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351

Wallace DC (2009) The pathophysiology of mitochondrial disease as modeled in the mouse. Genes Dev 23:1714–1736

Watkins-Chow DE, Pavan WJ (2008) Genomic copy number and expression variation within the C57BL/6 J inbred mouse strain. Genome Res 18:60–66

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A. 103:17600–17601

Wong K, Bumpstead S, Van Der Weyden L, Reinholdt LG, Wilming LG, Adams DJ, Keane TM (2012) Sequencing and characterization of the FVB/NJ mouse genome. Genome Biol 13:R72

Xia Y, Won S, Du X, Lin P, Ross C, La Vine D, Wiltshire S, Leiva G, Vidal SM, Whittle B, Goodnow CC, Koziol J, Moresco EM, Beutler B (2010) Bulk segregation mapping of mutations in closely related strains of mice. Genetics 186:1139–1146

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Migita S, Bonhomme F, Hjorth JP, Petras ML, Tagashira Y (1982) Origins of laboratory mice deduced from restriction patterns of mitochondrial DNA. Differentiation 22:222–226

Yu X, Wester-Rosenlöf L, Gimsa U, Holzhueter SA, Marques A, Jonas L, Hagenow K, Kunz M, Nizze H, Tiedge M, Holmdahl R, Ibrahim SM (2009) The mtDNA nt7778 G/T polymorphism affects autoimmune diseases and reproductive performance in the mouse. Hum Mol Genet 18:4689–4698

Zelnick CR, Burks DJ, Duncan CH (1987) A composite transposon 3' to the cow fetal globin gene binds a sequence specific factor. Nucleic Acids Res 15:10437–10453

Chapter 6
Epigenetic Control of Genome Expression

6.1 Introduction

From the standpoint of evolution, diploidy is generally considered advantageous for two reasons. First, because diploid organisms possess twice as many genes as haploids and in these conditions twice as many favorable mutations arise per generation. This of course increases the genetic diversity in the population and, finally, contributes to the progress of adaptive evolution. Diploidy is also considered advantageous because, when a recessive mutation occurs in a given gene, there is always a backup copy of the original allele on the other chromosome, offering a chance for the population to assess, with no risk, which one of the two alleles is most advantageous for the future of the species in a given environmental context. In most cases, the new mutant allele is neutral and has no selective advantage; sometimes it is harmful and is more or less rapidly eliminated. On rare occasions, it is beneficial and can then gradually replace the original allele.

However, mammals are not perfectly diploid since the males have an X and a Y chromosome while females have two X chromosomes. This difference, which is associated with sex determination, requires that a mechanism of gene dosage compensation be developed to equilibrate the transcriptional activity of the X-linked genes between the two sexes. Understanding this mechanism and its determinism has elicited a great number of investigations over the last fifty years, and scientists have discovered that, in eutherian mammals, the females are functionally haploid for the major part of their X chromosome. As we will explain in this chapter, this functional haploidy is controlled by an epigenetic mechanism leading to the inactivation of one of the two X chromosomes.

In the same mammals, scientists have also discovered that some autosomal regions, sometimes reduced to one or a few genes, are also functionally haploid, exclusively expressing the alleles inherited from one of the two parents and not those inherited from the other, due to the intervention of similar epigenetic mechanisms. These discoveries have broken the dogma of the superiority of diploidy over haploidy and have revealed the existence of a new kind of control of the transcriptional activity of mammalian genes. Although the epigenetic mechanisms

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_6

controlling the transcriptional activity of the X chromosome are not exactly the same as those at work for the autosomal regions, they nonetheless have so many similarities that we will describe them here, in the same chapter.

6.2 X-Chromosome Inactivation in Mammals

In mammals, the XX/XY sex-determination system is common, and only rare exceptions have been reported.[1] In the mouse, females have two large X chromosomes while males have an X and a Y on which the sex-determining region (*Sry*) is the master regulator of sex determination. In its absence, for example in mice with an XX or XO karyotype, the embryo develops as a normal, healthy and fertile female.[2]

The XX/XY system is both simple and robust, since relatively few anomalies in sex determination (intersexuality or sex ambiguities) have been reported, but the presence of two X chromosomes in the female versus a single one in males clearly raises a problem associated with gene dosage imbalance. For this reason, during their evolution mammals have found an efficient way to compensate (or more precisely to equilibrate) the transcription of X-linked genes between the two sexes.

6.2.1 In Female Mammals Only One X is Transcriptionally Active

The XX/XY sex-determination system exists in many diploid organisms, and different ways of solving the question of XX/XY dosage compensation have been retained. In the fruit fly *Drosophila melanogaster*, for example, the male-specific lethal (MSL) complex increases transcription of the single X chromosome to equalize expression of X-linked genes between the two sexes (Larschan et al. 2011). In *Cænorhabditis*

[1] At least two exceptions to the classical XX/XY mechanism of sex determination have been reported. The first one is found in wood lemmings (*Myopus schisticolor*), a species of Cricetidae rodent in which there are two types of X chromosomes (X and X*) and a Y chromosome. XX genotypes develop as females and XY develop as males, as in other mammals. However, both X*X and X*Y develop as females because the X* chromosome carries a mutation that inhibits the male-determining effect of the Y chromosome. The three categories of females (XX, X*X, and X*Y) are fertile, but X*Y females only produce X* ova. This sex determination system induces a strong distortion in the sex ratio (3/1 instead of the normal 1/1) and is considered an adaptation to the extreme seasonal reductions in population size that might otherwise threaten the survival of the species. Another remarkable exception is the mole vole *Ellobius lutescens*, another species of Cricetidae rodent in which both the male and female have the same odd number of chromosomes with a single X and no Y. In this species the sex-determination process is not yet completely understood.

[2] The development of testes as gonads also depends upon some other genes (*Foxl2, NrOb1, Sox9,* etc.).

elegans, dosage compensation is achieved by the female in which transcription from the two X chromosomes is simply halved (Kelly et al. 2002). In mammals, yet another solution has evolved with one of the two X chromosomes being inactivated in the female.

In 1961, the Harwell geneticist Mary F. Lyon suggested that, to ensure correct gene dosage compensation between male and female mammals, one out of the two X chromosomes was randomly and permanently inactivated during embryonic and adult life (Lyon 1961). This hypothesis, as reported by Lyon herself (Lyon 2002), was based on two main observations: (i) one X chromosome is sufficient for normal (female) mouse development; and (ii) mice heterozygous for some alleles at the X-linked coat color loci *mottled* (*Atp7a*^{Mo}) or *dappled* (*Atp7a*$^{Mo\text{-}dp}$) show a variegated effect in heterozygotes, with a pattern of mottling resembling that "seen in somatic mosaics". Lyon's hypothesis has been validated in a great number of mammalian species, and the mechanism of inactivation has been progressively elucidated at the molecular level.[3]

To explain how X-chromosome inactivation works, we have selected two examples: the first one is common and refers to tortoiseshell and calico female cats, while the second is historical and refers to glucose-6-phosphate dehydrogenase deficiency in women. The choice of these two examples may appear paradoxical in a book dedicated to the mouse, but it has the great advantage of being didactic.

6.2.1.1 Calico Cats and G6PD-Deficient Women

In the cat, the *Orange* locus is X-linked, it has two alleles: black O^b and orange O^o (or o), and no homolog on the Y chromosome. In male cats this locus determines two phenotypes: black (O^b/Y) or orange (O^o/Y), depending on the allele carried by the X chromosome. In females there are three genotypes: O^b/O^b, O^o/O^o, and O^b/O^o and also three phenotypes: black (O^b/O^b), orange (O^o/O^o), and a third phenotype called *tortoiseshell*, which is observed in the heterozygous O^b/O^o. This phenotype is clearly different from the uniform phenotype we would expect to get for a cat heterozygous for an autosomal gene involved in the determinism of coat color and exhibiting the classical dominant/recessive or semi-dominant allelic interactions. Here, in contrast, the phenotype suggests that the two alleles, O^o and O^b, are expressed independently and exclusively, rather than simultaneously in the pigment-forming cells (the melanocytes). In other words, the fur of each female cat appears as a mixture of hairs in which the individual melanocytes express either one or the other of two different alleles at the X-linked *Orange* locus. This is a clear-cut and classical example of the functional inactivation of one of the two X chromosomes in female mammals (Fig. 6.1).

[3] On July 22, 2011, at the occasion of the 50 Years of X-Inactivation Conference held in Oxford, the Lyon hypothesis became the Lyon law.

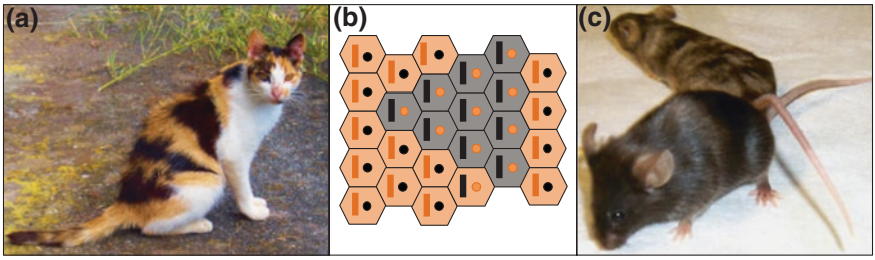


Fig. 6.1 *Calico cats and dappled mice*. **a** The figure represents a female cat with a typical "three-color" coat. Cats with such a coat color are called *calico* and are heterozygous for two different alleles at the X-linked *Orange* (*O*) locus: black O^b and orange O^o. The spots are either *black* or *orange* depending on the active X chromosome in the melanocytes. The *white* areas represent the unpigmented background and are due to a recessive autosomal spotting allele, called *piebald*. This allele, extremely common in the cat, makes the (*orange* or *black*) spots encoded by the O^b or O^o alleles even more visible (Courtesy of Dr. Abitbol, Alfort Veterinary School, France). **b** The diagram represents three contiguous clones of melanocytes, derived from independent stem cells in which a different X chromosome is inactivated. Since X inactivation occurs early in development and is irreversible, many of the observed spots in the adult cat represent a cluster of cells derived from the same stem cell. **c** The figure represents a female mouse heterozygous for the $Atp7a^{Mo\text{-}dp}$ (*dappled*) allele. Mutations at this X-linked locus are common and affect copper metabolism (Courtesy of Dr. Eppig, The Jackson Laboratory Bar Harbor, Maine, USA)

Looking at the fur of different female cats with a similar O^b/O^o genetic constitution, one may also note that the X chromosome that is inactivated in the melanocytes results from a random process because there is no specific pattern for the distribution of the orange or black pigment, while the proportion of orange/black fur remains close to 50 %. Also, it seems clear that once an X chromosome is inactivated, this status persists in the daughter cells, resulting in the appearance of a mosaic female made up of a mixture of cells, with one or the other X chromosome actively producing either one of the two alternative gene products at the *Orange* locus.[4] Since, as we shall discuss later, X inactivation occurs quite early in development, patches of cells with a similar pattern of X inactivation can become quite large and are easily seen on the female's coat. Some O^b/O^o female cats have an even more spectacular coat color pattern when, by chance, they also carry an autosomal spotting allele (for example *piebald*), because this allele makes the orange and black fur patches even more distinct on an otherwise white background. Female cats with this coat color pattern are called *calico*.

Another observation that illustrates well the consequences of X inactivation at the phenotypic level was published in 1962, by Ernst Beutler (Beutler et al. 1962),

[4] The term "mosaic" is appropriately used in this context (see Chap. 2) because all the cells in a female organism derive from the same egg and have the same genetic makeup at the *Orange* locus. The difference in gene (or allele) expression depends upon the active/inactive status of one of the two Xs. This results from an epigenetic mechanism, but not from a difference at the DNA or chromosome level.

a few months after the publication of Lyon's theory, and refers to the human genetic deficiency in glucose-6-phosphate dehydrogenase (G6PD). To explain their observation concerning the kinetics of dehydrogenation of glutathione (GSH) by the enzyme G6PD from the erythrocytes of heterozygous human females, Beutler and colleagues came to the conclusion that two populations of erythrocytes co-existed in females heterozygous for the X-linked deficiency (*G6PDX* gene) rather than a single one, as would have been the case for enzymes encoded by autosomal genes. Once more, the situation appeared to be the consequence of monoallelic and independent expression of G6PD in the individual red cells of the heterozygous patients.

Many more examples of mosaicism have been reported in female mammals, including humans, to illustrate this point. The so-called Barr body, which was observed and reported years ago, even before Lyon's hypothesis, as a darkly stained dot in the nucleus of cells prepared from oral swabs, represents a heteropycnotic X chromosome. Karyotypes with X-chromosome aneuploidy (monosomics or XO, trisomics XXX or XXY male patients) display a number of Barr bodies that always contain one less than the total number of X chromosomes in the karyotype, indicating that there is a biological mechanism that somehow "counts" the total number of X chromosomes in mammalian cells in addition to the mechanism inducing inactivation of all X chromosomes but one.

6.2.1.2 X Inactivation is a Two-Step Process that Occurs Early During Embryonic Life

The mechanism and precise timing of X-chromosome inactivation (XCI) were debated for a few years after the initial publications of Lyon's hypothesis. Nowadays it is established that, in the mouse, two different forms of X-chromosome inactivation occur successively during early female embryogenesis. The first is an *imprinted* or selective inactivation, which starts at the 4–8-cell (morula) stage and affects only the paternal X chromosome (X_p).[5] By embryonic day 6.5 (i.e., when gastrulation begins), the X_p is reactivated in the cells that will give rise to the embryo proper, then the classical X_p or X_m *random* inactivation ensues that will be retained as such for the rest of the organism's life (Morey and Avner 2011; Pollex and Heard 2012).[6] In contrast to the tissues of the embryo proper, the imprinted or selective X_p inactivation is maintained in the extra-embryonic tissues (placenta) for the rest of gestation.

[5] This X_p-specific inactivation is consistent with the observation that, at the pachytene step of male meiosis, the X_p is condensed with the Y chromosome in an inactive XY body while, at the same pachytene stage of female meiosis, the two X chromosomes are visible and form a normal bivalent.

[6] Unlike in eutherian mammals, the imprinted X_p inactivation persists in all cells of protherian mammals (marsupials) including in the cells of the embryo proper.

As we already mentioned, the randomness of X-inactivation in the embryonic and adult tissues is faithfully translated at the phenotypic level. For example, when looking at the external appearance of adult calico cats one can observe that in most instances approximately 50 % of their spots are orange (O^o/–) while the other 50 % are black (−/O^b). This randomness was also demonstrated more rigorously by Davidson and colleagues (Davidson et al. 1963), who derived clones of epithelial cells from female patients heterozygous for two different forms of the enzyme glucose-6-phosphate dehydrogenase ($G6PD^A$/$G6PD^B$). In their experiment, the authors found that of 14 clones of cells derived from the same heterozygous patient, seven showed the *A* form of the enzyme while the other seven showed the *B* form, and none contained both the *A* and *B* forms.

6.2.1.3 X-Inactivation Is Complete … or Nearly so

X inactivation is thought to be highly stable in somatic cells and does not revert in the cells of the developing embryos, after implantation or in the cells of adult females. However, it has been reported that a few genes on the inactivated X chromosome could reactivate at low levels during aging. This is obviously a consequence of some relapse in the X-inactivation process, but remains marginal and concerns only a minority of the X-inactivated genes.

The situation is rather different with another limited set of genes, which are on the mouse X chromosome and escape X inactivation completely. Most of these "escapees" map to the pseudo-autosomal region (PAR), which means that they have a homolog on the Y chromosome. The pseudo-autosomal regions on the X and Y chromosomes pair and recombine during meiosis, (almost) as if they were autosomal, and it makes sense to believe that this is probably the reason why they are not inactivated: after all, there is no reason to apply any form of dosage compensation to these genes. *Steroid sulfatase* (*Sts*) is the best known example of these genes mapping to the PAR; mice homozygous for a deficient allele (Sts^-/Sts^-) have been reported as a model for a common neurodevelopmental disorder in humans, *attention deficit hyperactivity disorder* (ADHD) (Trent et al. 2011). However, unexpectedly, the same Sts^-/Sts^- mice are not a model for the human X-linked recessive disease *ichthyosis,* although human patients appear to be affected on the orthologous gene *STS*.

Besides the genes mapping to the PAR, some other genes mapping to the X chromosome also escape inactivation and are found to be transcribed from the inactive X chromosome. Most of these genes have (or had) a homolog on the Y chromosome or elsewhere in the genome, but this homolog is no longer functional. They are orphan genes and probably do not encode any functional proteins. The reason why these genes escape inactivation is unclear, but this does not seem to be a problem since XO females appear to be normal though sub-fertile. In contrast, in humans, where many more genes escape X inactivation, XO females present a severe phenotype known as Turner syndrome, which is probably due to these escapee genes, both in the PAR and elsewhere on the X chromosome.

6.2.2 The Mechanisms Controlling X-Chromosome Inactivation

6.2.2.1 Characterization of an X-Inactivation Center (XIC)

Elucidating the mechanisms leading to X-chromosome inactivation consists of understanding how two genetically identical and transcriptionally active X chromosomes, that lie within the same nucleus, can be differentially treated in such a way that one of them remains active while the other is inactivated. The first important observation in this matter was that all inactivated genes are on the same chromosome, while all active alleles are on the other. To interpret this observation, scientists hypothesized that the inactivation was essentially a chromosomal issue and that a master switch, controlling inactivation, might exist somewhere on the X chromosome from which inactivation starts and spreads along the rest of the chromosome. The identification of this master switch or *inactivation center* (XIC) was achieved in several laboratories in the mid-1990s, using strategies that are common in genetics, consisting of the demonstration of the physical association of a short chromosomal segment with its potential to inactivate the flanking regions of a given chromosome. Studying the consequences of various reciprocal translocations and deletions involving the X chromosome and mouse autosomes allowed the demarcation of a region of chromosome X with these properties. Confirmation of these observations came from experiments of transgenesis with cloned DNAs of various size followed by analysis of the consequences of the transgene on the flanking regions.

The extent of the region enclosing the inactivation center has been defined by studying X-chromosome deletions and by performing transgenesis in embryonic stem cells. Both experiments have permitted the characterization and delimitation of a region spanning a few hundred kilobases. Female embryonic stem cells (ES cells) have been invaluable tools for studying the XIC because these cells have their two X chromosomes active when undifferentiated, while X inactivation proceeds, as in embryos, when they start to differentiate in vitro.

A second discovery was that the XIC contains a gene, called *Xist* (for *X-inactive-specific-transcript*) that is transcribed into a non-coding RNA expressed only from the inactive X chromosome (Brown 1991). The *Xist* RNA was found to coat the inactive X chromosome in *cis* and to correlate with the onset of X inactivation.

Although the properties and localization of *Xist* RNA strongly suggest that it should be a key element in the X-inactivation process, further experimental evidence was required to show that this locus is necessary for inactivation of the X chromosome. This was shown through the use of various deletions of the *Xist* gene that prevent production of full-sized *Xist* RNA (Penny et al. 1996; Marahrens et al. 1997). In these cases the chromosome bearing the deletion is not inactivated, indicating that a complete, intact *Xist* gene is required, in *cis*, for inactivation to take place.

Further support for a critical role of XIST comes from experiments in which the *Xist*-cDNA, under an inducible promoter, was inserted into an autosome in male ES cells. Induction of *Xist*-RNA provoked coating of the chromosome in *cis* and repression of gene transcription for this autosome. Although other factors are probably also involved, these experiments demonstrated that *Xist*-RNA is a key trigger for chromosome-wide silencing and that it may do so by binding to the chromosome from which it is expressed. These experiments also demonstrated, along with previous studies from X-autosome translocations, that specific X-linked sequences are not required for *Xist*-RNA to coat a chromosome.

The XIC candidate region harbors four non-protein-coding genes, *Xist*, *Tsix*, *Jpx,* and *Ftx*, which are involved in X-inactivation. The XIC also contains binding sites for both known and unknown regulatory proteins.

The *Xist* transcript has no significant open reading frame and the product remains in the nucleus, coating the inactive X chromosome. This suggests that *Xist* is among those loci that produce a functional RNA molecule that is never translated into a protein (a non-coding RNA—see Chap. 5). *Xist* expression is detected early in pre-implantation development, often from both X chromosomes, just prior to X inactivation at the 4–8-cell stage (Okamoto et al. 2004; Patrat et al. 2009). In the mouse, the paternal X chromosome is initially subject to X inactivation as a result of an imprint in the gametes that leads to the paternal nonrandom inactivation found in extra-embryonic cells. Later, in the inner cell mass, *Xist* is activated from one of the two X chromosomes in cells that will form the epiblast. Random X inactivation follows and *Xist* transcription on the active X chromosome is silenced. Recent evidence suggests that XIST regulation involves a combination of *cis*-elements including antisense transcription as well as *trans*-acting factors that are tightly integrated with the pluripotent and stem cell proteins (for a recent review, see Augui et al. 2011).

The inactive X chromosome has been associated with several putative epigenetic marks (or non-sequence-based heritable changes) including DNA methylation, histone modifications, and Polycomb group complexes. DNA methylation is probably the best studied to date. Methylation of the cytosine base occurs enzymatically after DNA synthesis, and in mammals is restricted to the dinucleotide 5′-C_pG-3′ (C_pG). About 7 % of C_pGs are present at relatively high density in clusters called C_pG islands, which are usually located at the 5′ ends of genes. The remaining C_pGs are dispersed throughout the genome, usually as singlets. Most C_pG islands are unmethylated, but those near inactivated genes on the X chromosome, and those near some imprinted genes on autosomes, are methylated. Methylated C_pG islands repress transcription, and most silent genes on the inactive X chromosome have such methylated C_pG islands in normal cells. It is believed that DNA methylation acts in a synergistic way with other chromatin modifications to lock in the inactive state in a highly stable fashion in somatic cells.

The mouse has played a fundamental role in our understanding of the mechanisms of gene regulation and expression underlying processes such as X inactivation, as it has rendered observations and experiments possible that were not possible in any other species up until quite recently.

6.2.2.2 X-Inactivation Skewing

Most women heterozygous for the X-linked mutation DMD (Duchenne muscular dystrophy—*DMD*$^{+}$/*DMD*^{mut}) remain completely asymptomatic during their life and are generally unaware that they are carriers until they give birth to an affected son. This situation is common to many other pathologies where females are heterozygous for a deleterious X-linked mutation. The lack of overt phenotype or only mild phenotype in females is generally explained by considering that around 50% of their cells express the normal allele from the active (transcribed) X chromosome, with the mutated allele being on the silent, inactive X chromosome. This explains why these carrier females are protected from the clinical effects of X-linked mutations such as in the case of the *DMD* gene.

However, such situations of intercellular complementation are far from being the rule, and after careful analysis of other X-linked human pathologies, it has been observed that X inactivation may occur randomly at first (i.e., 50 % X^{+}/50 % X^{mut}) but, with time, the cells in which the X chromosome carries an allele with deleterious effects (X^{mut}) are counter-selected more or less efficiently, depending on the case, giving the impression of X-inactivation skewing. This is the case in a form of X-linked mental retardation (XLMR), ATR-X syndrome, which is caused by mutations in a ubiquitously expressed, chromatin-associated protein and in which phenotypically normal female carriers have highly skewed X-chromosome inactivation of the X chromosome that carries the mutant allele. Interestingly, the homologous disease has been modeled in mice heterozygous for a null *Atrx* orthologous allele, and it has been observed that X-chromosome inactivation is balanced early in embryogenesis but becomes skewed over the course of development because of a strong selection favoring cells expressing the *Atrx* wild-type allele (Garrick et al. 2006).

Selection against the cell lineage that carries the mutant allele on the active X chromosome appears logical, especially if it is the price to pay for surviving in better conditions, but it is not the rule. For example, unfavorable skewing of X inactivation has been reported in young females suffering from hemophilia B where the paternal X chromosome, carrying a normal copy of the FIX gene, was predominantly the inactive one, leading to the phenotypic expression of hemophilia B in these young girls (Espinós et al. 2000).

X-inactivation skewing is sometimes influenced by chromosomal rearrangements. An excellent example of such skewed X inactivation is provided by the *T(X;16)16H* (or Searle's) reciprocal translocation in the mouse. In this translocation, a piece of the telomeric region of chromosome X is attached to the centromeric part of chromosome 16, and vice versa. As expected, the piece of X chromosome that carries the X-inactivation center is inactivated, but inactivation spreads over the breakpoint and concerns all the genes on the piece of chromosome 16 that is attached to the broken X, resulting in a deleterious functional haploidy. Conversely, all the X-linked genes on the non-inactivated piece of X chromosome are expressed, where they should not be. In fact, for the female mice heterozygous for Searle's translocation, the only way to survive is to

inactivate their normal X chromosome. It is likely that such a situation, which is extreme in the case of *T16H* mice, probably exists with other mutations, although sometimes with a less dramatic effect—leading to a less extreme skewing.

At this point, it is interesting to note that, 40 years ago, Cattanach had already reported skewed X inactivation in F1 hybrid mice. He observed that, depending on the cross and strains involved, the percentage of inactivated X chromosomes was different for X chromosomes of different genetic origins. Cattanach quantified his observations by defining four alleles at a locus that he designated as the X-inactivation controlling element (symbol *Xce*) with four alleles: $Xce^a < Xce^b < Xce^c < Xce^d$ in order of the tendency of the X chromosome to remain active (Simmler et al. 1993; Thorvaldsen et al. 2012). As of today, the identity of the *Xce* locus remains unknown, although its genetic localization has been much refined, indicating that multiple elements on the X chromosome contribute to the *Xce* and that some of these may lie within the X-inactivation center (see below).

6.3 Parental Imprinting of Autosomal Genes

As discussed above, X-chromosome inactivation is an original and sophisticated method which has evolved to equilibrate the transcriptional activity of the genes on this chromosome between male and female mammals. Being epigenetic by nature, X-chromosome inactivation does not alter the DNA sequence and is completely erased when the primary germ cells enter gametogenesis. However, the X chromosome is not the only segment of the mammalian genome that can be modified epigenetically, as we will now discuss.

6.3.1 Evidence of Genomic Imprinting in the Mouse

6.3.1.1 The Unusual Behavior of the Hairpin-Tail Allele at the *T*-Locus

The *T*-locus of the mouse (brachyury *T*-Chr 17) has several mutant alleles; some are dominant while others, mostly found in wild mice, are recessive. Dominant alleles have an effect on the notochord derivatives and are characterized, when heterozygous, by a shortened tail with extensive variation in expressivity. *T/T* homozygotes die during embryonic development, at about mid-gestation.

Hairpin-tail (T^{hp}) is unique in the allelic series at the *T* locus, in the sense that the phenotype of the heterozygote offspring depends upon the origin of the mutant allele. When T^{hp} is inherited from a $T^{hp}/+$ male mated to a wild-type female (+/+), the offspring are all viable and about 50 % of them have a shortened tail, as expected. However, when the cross is set up the other way around (i.e., between a

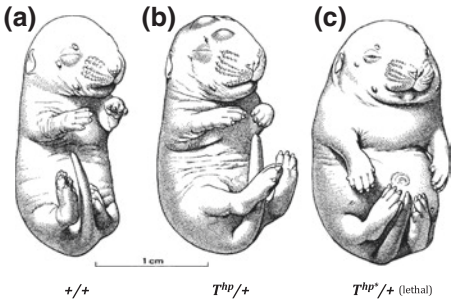


Fig. 6.2 *Inheritance of the hairpin-tail mutation*. 17-day-old embryos collected in the uterus of a T^{hp*}/+ mother mated with a T^{hp}/+ male: +/+; T^{hp}/+; T^{hp*}/+. Embryo (**a**) is normal. In embryo (**b**), the T^{hp} allele is of paternal origin, while it is of maternal origin* in embryo (**c**). Offspring with a genotype $T^{hp}*/T^{hp}$ die shortly after implantation and are not represented in the figure. The tail shortening effect of T^{hp} is not obvious in this figure, especially in embryo (**b**), probably for lack of expressivity (from Johnson 1974a). T^{hp} has been characterized as a large deletion on chromosome 17, which includes the (imprinted) gene encoding IGF2R

wild type +/+ male and a T^{hp}/+ mutant female), the progenies are reduced (nearly halved) and no mutant phenotypes are observed: they all die in utero at a relatively late stage of gestation.[7] This peculiarity of the hairpin-tail allele, which was first reported in 1974 as "*a case of post-reductional gene action in the mouse egg*" (Johnson 1974a, b), is not a simple maternal effect, since T^{hp}/+ × T^{hp}/+ matings produce two types of T^{hp}/+ heterozygous embryos: one is viable ab utero with a short tail, while the other is unviable (Fig. 6.2).

Nowadays, we know that the T^{hp} allele is associated with a deletion in the centromeric region of chromosome 17 (T-associated maternal effect—*Tme*). We will later discuss the molecular nature of this structural change and its consequences, but at this stage and from a historical point of view, it is important to note that the identification of this allele at the *T* locus was quite fortunate. If, by chance, the original T^{hp} mutant allele had occurred in a female germ cell it would have been lost and the discovery of a "post-reductional gene action" would have been delayed.

6.3.1.2 The Fate of Embryos Resulting from the in Vitro Re-Association of Pronuclei

For many years, and for technical reasons, it was impossible to grow mouse eggs in vitro, from the one-cell stage up to the stage of expanded blastocyst. Once this difficulty was overcome, one of the first experiments undertaken by embryologists was to try and reconstruct artificially diploid embryos by re-associating pronuclei

[7] Some exceptions have been reported, but they are extremely rare and fall well below the expected 50 %.

from embryos at the one-cell stage in different combinations. The rationale for undertaking this sort of experiment was to check whether a given haploid genome could merge with any other haploid genome to result in a viable mouse organism. Such experiments were completed in the early 1980s, in particular in England and in the USA, and led to the unambiguous conclusion that the development to term of reconstructed pseudo-diploid embryos requires the association of a maternally derived and a paternally derived pronucleus. Any other association (i.e., two male pronuclei or two female pronuclei) appeared lethal a few days after implantation (Barton et al. 1984; McGrath and Solter 1984; Surani et al. 1984) (Fig. 6.3).

The result of these experiments suggested that the haploid genome in a pronucleus was marked in a specific manner according to its parental origin, and that the male and female contributions were not functionally equivalent. This mark has become known as the *parental genomic imprint* or simply *genomic imprinting*.

Other experiments, focusing on the study of the developmental potentialities of cells derived from either gynogenetic embryos (resulting from the association of two female pronuclei) or androgenetic embryos (resulting from the association of two male pronuclei), merged together or associated independently with cells of a normal embryo in a single chimeric organism, indicated that androgenetic cells preferentially contribute to the formation of extra-embryonic tissues while gynogenetic cells, in contrast, preferentially contribute to the formation of embryonic

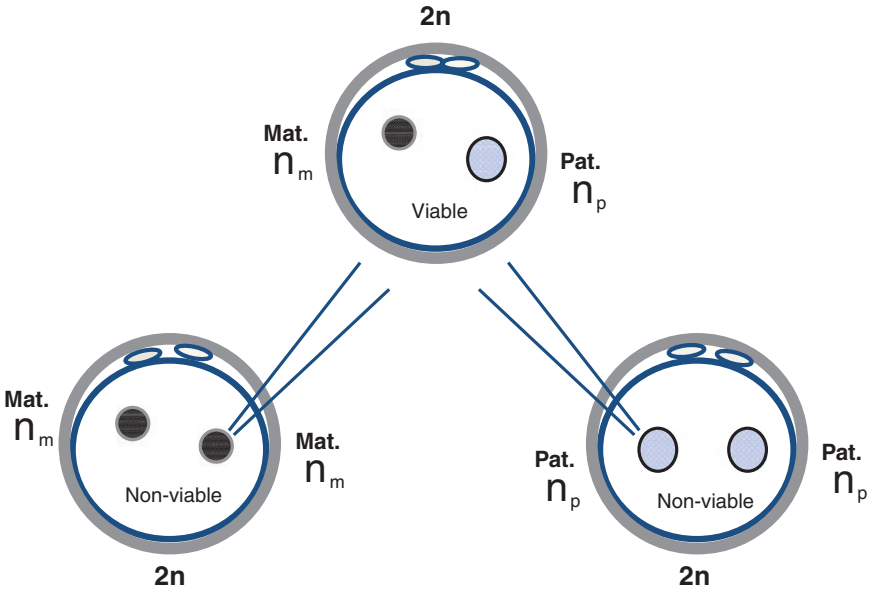


Fig. 6.3 *The fate of reconstructed, pseudo-diploid embryos*. The development to term of reconstructed, pseudo-diploid embryos requires the association of maternally derived and paternally-derived pronuclei. Reconstructed embryos with either two maternal or two paternal haploid sets are unviable

tissues. Another conclusion that can be drawn from these experiments is that parthenogenetic development is strongly hindered in the mouse although it occurs, occasionally, in other vertebrate species (it is common in fish and some reptiles, and has also been reported in birds).

6.3.1.3 The Fate of Embryos Resulting from Uniparental Disomies

The conclusions of the experiments reported above have been confirmed and refined by another totally different kind of experiments, achieved mostly in England, in the mid-80 s at the Harwell MRC Laboratory, by B.M. Cattanach, C.V. Beechey, J. Peters, and A.G. Searle. These experiments made use of two types of chromosomal rearrangements (Robertsonian translocations and reciprocal translocations) that were available in the large genetic repository at Harwell.

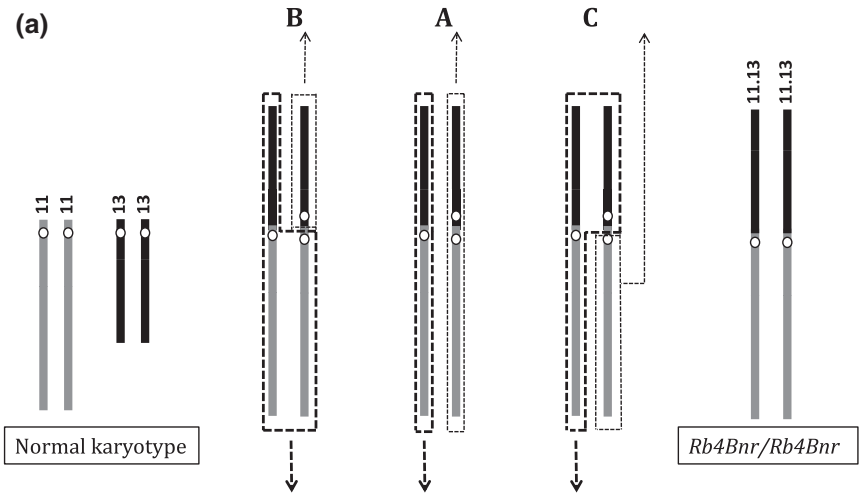
As described in Chap. 3, devoted to cytogenetics, mice whose genetic constitution consists of a single Robertsonian translocation plus the two acrocentric chromosomes whose arms are homologous to the arms of the Robertsonian translocation are perfectly normal since they have a balanced karyotype although reduced by one centromere. Such mice, however, often produce a high percentage of unbalanced gametes—i.e., gametes with either one extra (acrocentric) chromosome or, reciprocally, with one missing (acrocentric) chromosome. As we already discussed, these unbalanced gametes, resulting from meiotic non-disjunction, yield trisomic or monosomic embryos when merging with a normal gamete (Fig. 6.4).

In the mouse, most trisomic and all monosomic embryos die in utero at a stage of development that varies with the chromosome involved.[8] However, when by chance an unbalanced gamete with, for example, one missing acrocentric chromosome combines with a gamete with one extra chromosome of the same pair, this results in an embryo with a $[(n-1)+(n+1)] = 2n$ (euploid) chromosome complement, regardless of whether the two chromosomes of the pair in question were contributed by one and the same parent or not. Such embryos, with the two chromosomes of a given pair originating from the same parent, are said to result from uniparental disomies (UpDi).[9]

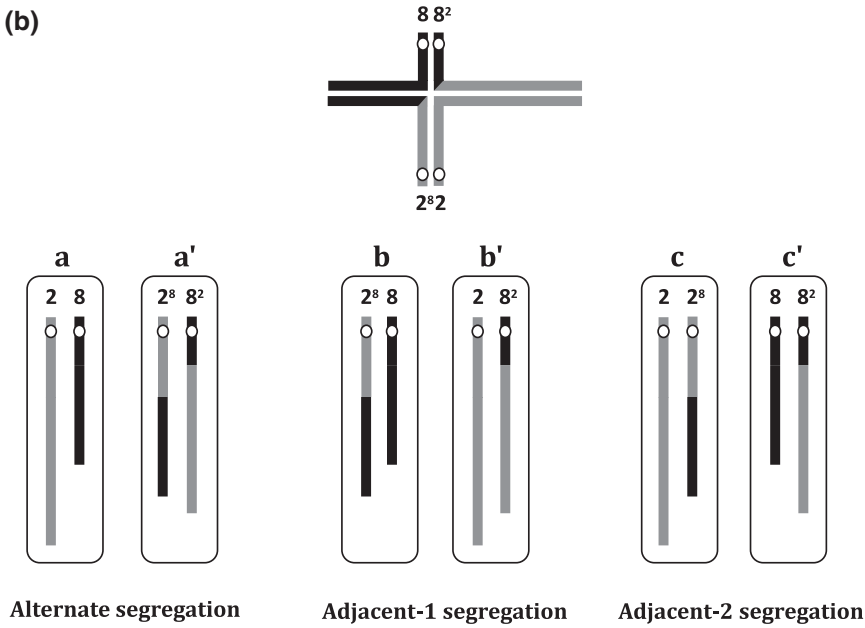
The observations by Cattanach and colleagues, made on the progenies of mice with a variety of different chromosomal translocations, were that viable and normal embryos resulting from complementary double non-disjunctions (UpDis) were (i) rather rare and (ii) very much dependent on the chromosome pair involved. In fact, in many instances, dramatic effects on development, including enhanced or retarded growth and sometimes lethality in utero, could be observed in the progenies (Cattanach and Kirk 1985; Cattanach 1986). Cattanach demonstrated that only a few chromosomes could be inherited as uniparental disomies, still leading

[8] *Ts19* is the only trisomy viable ab utero but only a few mice survive after 10 days.

[9] Uniparental disomies can be of maternal (MatUpDi) or paternal (PatUpDi) origin.



Meiosis in a Mouse heterozygous for a Robertsonian translocation



Meiosis in a Mouse heterozygous for a reciprocal translocation

Fig. 6.4 *Double non-disjunction in mice heterozygous for translocations*. In mice heterozygous for translocation the meiotic process often results in the production of a high percentage of aneuploid gametes due to the abnormal segregation of the chromosomes. **a** Represents the disjunction of chromosomes 11 or 13 in mice heterozygous for the Robertsonian translocation *Rb(11.13)4Bnr*. When a gamete with an extra chromosome arm merges with a normal gamete, this results in a trisomic embryo (see Chap. 3 for explanations). However, when the same aneuploid gamete merges with a complementary unbalanced aneuploid gamete missing the same chromosome arm, this recreate a normal (2n) karyotype with the exception that, in this case, the same parent provides the two chromosomes of a given pair and the other parent none of the gamete of the pair in question. In this case, the embryo is said to result from uniparental disomy (UpDi). Such embryos are viable only when the two elements of a chromosome pair involved in the UpDi are not imprinted. In the original experiments by Cattanach and colleagues (see text), identification of the parental origin of the chromosomes was done by using the phenotypic genetic markers vestigial tail (*vt*) for chromosome 13 and dominant, wavy coat (Re^{wc}) for chromosome 11. Nowadays, molecular markers like SNPs or microsatellites would rapidly distinguish the origin of the different chromosomes in such a cross. In the cross represented here, Cattanach and colleagues observed that the offspring resulting from MatUpDi11 (maternal uniparental disomy of chromosome 11) were smaller than their normal sibs while the offspring resulting from PatUpDi11 were bigger. This was a demonstration of the parental imprinting of (at least a segment of) the chromosome 11. Doted lines show the three different segregations of chromosome 11 including non-disjunctions. (Adapted from Cattanach's original drawings). **b** Represents the disjunction of chromosomes 2 and 8 in mice heterozygous for the reciprocal translocation *T(2;8)26H*. These mice produce a variety of gametes (**a–a′**, **b–b′**, **c–c′**) with a variety of chromosomal segment association, depending on the type of segregation (see Chap. 3 for explanations). Some of these gametes carry duplicated segments (for example, **b′** and **c** for Chr 2; **b** and **c′** for Chr 8) while some others carry segmental deletions (for example, **b** and **c′** for Chr 2; **b′** and **c** for Chr 8). The gametes with either a segmental deletion or a segmental duplication (**b–b′** and **c–c′** on the picture) produce unviable offspring when they merge with a normal gamete—only gametes of the **a–a′** type produce embryos with a balanced (viable) karyotype. When the cross is between two progenitors heterozygous for the same reciprocal translocation, there are rare cases where two gametes resulting from complementary non-disjunctions fuse together, restoring a balanced karyotype (i.e., when the duplications complement the deficiencies). These offspring, resulting from complementary uniparental partial disomies, are rare but they can be identified if genetic (or molecular) markers segregate in the cross, labeling the various chromosome arms. A major impediment to this kind of experiment is that many reciprocal translocations, when heterozygous, are sterile in one sex or the other. The production of neonates resulting from complementary double non-disjunction is also laborious because, unlike for Robertsonian translocations, mice heterozygous for reciprocal translocations produce small-sized progenies due to embryonic lethality (semi-sterility; see Chap. 3 for explanations). (*Adapted from Cattanach's original drawings*)

to normal healthy offspring. In all other cases, anomalies were observed, generally associated with difference in body size.

The general conclusions of these experiments are that normal development to term of a mouse embryo requires that some specific chromosomes be inherited from the mother or from the father, and sometimes from both the father and the mother (for example, chromosomes 7 or 11). This again suggested that a parent-of-origin-specific expression exists, at least for some genes, and for one and/or the other of the two parental chromosome homologs.

In addition to this series of experiments (made with mice heterozygous for Robertsonian translocations and concerning intact, complete acrocentric chromosomes), scientists at Harwell used another approach to screen the whole mouse

genome for specific imprinted regions. The strategy made use of an assortment of reciprocal translocations, a very common type of chromosomal rearrangement, resulting from the reciprocal exchange of chromosome arms between two non-homologous chromosomes. Here again, mice heterozygous for reciprocal translocations produce a variety of aneuploid gametes and, by inter-crossing such mice, it is possible to obtain normal, 2n embryos whose genomes result from the fusion of complementary unbalanced gametes. These experiments were arduous and required many crosses because, as we explained in Chap. 3, the progenies of mice heterozygous for a reciprocal translocation are much reduced in number. After carefully screening hundreds of progenies, the scientists at Harwell could observe the presence (or suspect the absence) of conceptuses resulting from uniparental duplication/deficiency for a particular chromosomal region and, finally, they could summarize their observations by drawing a chromosomal map indicating the maternally or paternally imprinted chromosomal regions (See Fig. 6.5).

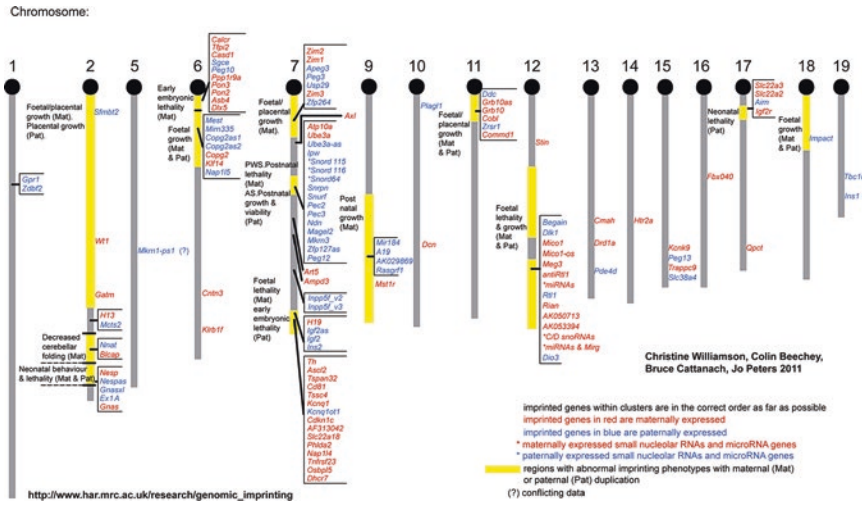


Fig. 6.5 *The Harwell map of mouse imprinted genes and regions*. Some chromosomal segments (outlined on the map) must be inherited from the male parent or from the female parent or, sometimes, simultaneously from both the male and the female parents. This is a consequence of genomic imprinting, which occurs during the process of gamete formation, and results in the functional inactivation of some specific genes encoding proteins or RNAs. The size of the imprinted segments has been estimated based on experimental data (see references), and in most instances it is excessively large compared to the actual size of the cluster of imprinted genes (1 Mb on average). Most (although not all) imprinted genes in the mouse are also imprinted in human and rat species. The establishment of this map has required an enormous investment in terms of crosses, and was possible only in a few laboratories (like MRC Harwell) where a large repository of translocations of all kinds existed. This map is now being progressively refined by direct analysis of the transcripts

6.3.2 Characterization of the Imprinted Regions in the Mouse

6.3.2.1 Imprinted Regions Harbor Genes that are Transcribed Exclusively from One Allele

The first imprinted region that was (partially) characterized at the molecular level was precisely the one that was discovered first and is associated with the "hairpin-tail phenotype". The characterization of the region in question was achieved by making a fine genetic map of the chromosome 17 proximal segment and performing a quantitative assessment of the transcription products of the genes mapping to that region. Providentially, another allele at the same *T/t* locus (t^{Lub2}) was discovered, which is recessive and associated with similar developmental defects as T^{hp}. When the chromosome carrying the t^{Lub2} mutation is inherited from the mother, embryos heterozygous for this mutation are severely affected by edema and death generally occurs between days 15–17 of gestation, just as for $T^{hp}/+$ mice born to a $T^{hp}/+$ mother (Winking and Silver 1984). Genetic and molecular analyses indicated that T^{hp} and t^{Lub2} were overlapping deletions of chromosome 17, with T^{hp} spanning a distance of about ~7 Mb and t^{Lub2} only ~0.8 Mb.

The t^{Lub2} haplotype has been characterized in detail, and several genes (Chr 17 cen—*Plg*, *Igf2r*, *Tcp1*, *Sod2*) have been identified within the deleted region. Remarkably, among all these genes *Igf2r*, the gene encoding the insulin-like growth factor type-2 receptor (IGF2R) appeared to be transcribed exclusively from the maternal allele, while the other genes were transcribed from both the paternal and maternal alleles.

Considered together, these observations explain all the observed phenotypes; in short, since *Igf2r* is deleted in the T^{hp} and t^{Lub2} chromosomes, and given that *Igf2r* is not transcribed from the paternal allele, any embryo with a $T^{hpM}/+\ ^{P}$ or $t^{Lub2M}/+\ ^{P}$ constitution has no functional IGF2R and accordingly cannot survive to birth. Embryos with the reciprocal genotype (i.e., $T^{hpP}/+\ ^{M}$ or $t^{Lub2P}/+\ ^{M}$) are normal since the maternal copy is intact and transcribed, exactly as in normal embryos. For all other genes, hemizygous embryos survive normally as they generally do with most other autosomal genes (Barlow et al. 1991).

Igf2r encodes a trans-membrane receptor protein whose function is to transport mannose-6-phosphate tagged proteins and insulin-like growth factor 2 (IGF2) to lysosomes; it is an essential protein for the completion of a normal gestation. The conclusions drawn from these observations have been validated by studying, independently, the fate of embryos inheriting a non-functional copy (i.e., a knockout allele—see Chap. 8) of the *Igf2r* gene from their mother or from their father.

In a series of experiments performed two years later, i.e., once the detailed mechanisms generating imprinting were unraveled, scientists created a non-imprinted allele of *Igf2r* (designated *R2Delta*) by deleting an essential element repressing the paternal allele in mouse ES cells (actually the ICE—see below). Maternal inheritance of this *R2Delta* allele had no phenotype, as expected. However, paternal inheritance resulted in biallelic expression of *Igf2r*. In this case,

embryos were affected by a 20 % reduction in body weight late in embryonic development that persisted to adulthood. Paternal inheritance of the functional *R2Delta* allele rescued the lethality of a maternally inherited *Igf2r* null allele and a maternally inherited *Tme* (T-associated maternal effect) mutation. These data suggested that one of the biological reasons for imprinting *Igf2r* is probably to trigger an increase in body weight at birth. These data confirmed the importance of the *Igf2r* gene in the *Tme* deletion phenotype (Wutz et al. 2001).

The second region that was recognized as imprinted, and characterized at the molecular level, was the telomeric region of chromosome 7. This region was identified by studying the progeny of an intercross between mice heterozygous for the reciprocal translocation *T(7;18)50H*. Embryos with the maternal duplication and paternal deficiency of distal Chr 7 (*MatDp7/PatDf7*) are growth-retarded and die around day 16 of gestation; the reciprocal maternal deficiency and paternal duplication embryos (*MatDf7/PatDp7*) die at an unidentified but much earlier stage. The imprinted region harbors, among others, the gene encoding insulin-like growth factor 2 gene (*Igf2*), a gene functionally and physiologically related to the gene encoding its receptor *Igf2r* (DeChiara et al. 1991).

IGF2 is a growth-promoting hormone acting during gestation and sharing structural similarities with insulin. *Igf2* is imprinted differently from *Igf2r* since it is transcribed exclusively from the paternal allele. The observations relative to the growth retardation of the embryos resulting from chromosome 7 uniparental disomies have been confirmed by studying the mice carrying a null (knockout) allele of *Igf2*. As expected, non-complementation of the Igf^- allele by the normal *Igf2* allele was observed when the wild-type allele was inherited from the mother.

6.3.2.2 Making the Inventory of Imprinted Genes in the Mouse

Many genes have been progressively discovered in the various imprinted regions identified by Harwell's scientists, and a good proportion of these regions have now been characterized at the molecular level. As indicated on the map (Fig. 6.5), there are at least 15 and probably up to 25 imprinted regions spread over 16 different autosomes and these regions are apparently distributed randomly, i.e., with no specific pattern. They are either telomeric or centromeric and harbor clusters of genes (from 3 to 11) rather than single independent genes. Some geneticists think that this clustering of the imprinted genes is probably not by chance, and may reflect subordination to a common mechanism of inactivation. This conclusion, however, should be reconsidered when a greater number of imprinted genes or regions are identified in different mammalian species.

The genes mapping to the same imprinted cluster do not appear to be functionally related. Even more surprisingly, some genes in a given cluster are maternally expressed while others are paternally expressed (for example, *Igf2* and *H19* on distal Chr 7). This is in good agreement with the original observation at Harwell that,

at least for some pairs of chromosomes, one element must be inherited from the father and the other from the mother.

As we mentioned, the function of the genes mapping to the imprinted clusters is not always fully characterized, and for some of them it may take some time before we precisely determine all their functions. This is particularly true if we consider that some of these genes, for example *H19*, do not encode proteins but non-coding RNAs instead.

The analysis of the transmission of knockout (null) alleles, produced by in vitro gene targeting in the mouse, will be of great help for the future identification of imprinted regions or genes. It seems, however, that genes of this category represent only a minority of the genes because if the wild-type alleles of the genes that have been knocked out were imprinted, their uniparental transmission to the progeny would be impossible or associated with some pathology, and this would almost certainly have already been noticed by researchers. The analysis of the transmission patterns of knockout alleles is indeed an efficient way to screen for genomic imprinting in the mouse, and the occurrence of any phenotypic alterations exclusively transmitted by one sex and not by the other should trigger curiosity and call for further investigation. Similarly, identification of a new imprinted gene in humans (or any other mammalian species) should be considered as an indication for a candidate in the homologous region in the mouse.

As of today, the number of imprinted genes reported in the mouse is around 140. Studies of the total number of imprinted genes are currently being refined by other methods (Yu et al. 2012). Sequencing the whole transcriptomes of interspecific mouse hybrids resulting from crosses in both directions (for example, a female of a laboratory inbred strain × a *Mus m. musculus* male or vice versa), and looking for tissue/cellular distribution of species-specific SNPs is a promising way of achieving the complete inventory of imprinted genes in the mouse (Fig. 6.6).

Of the 140 genes that have been reported as being imprinted in the mouse, a quite large proportion has also been found to be imprinted in humans, but exceptions exist. *Igf2,* for example, has been found to be imprinted in the human, rat, and mouse species but the gene encoding the receptor for this molecule, *Igf2r*, is imprinted in the rat and mouse species but not in humans (Weidman et al. 2006). In addition to this observation, it is worth noting that, from interspecific comparisons that have been made, it seems that the degree of homology in terms of imprinted genes parallels the phylogenetic distances. This is not so surprising and, with a better knowledge of the imprinted genes across mammalian species, it should be possible to learn more about their function. Already, by comparing the known functions of the imprinted genes in the three above-mentioned species (human, mouse, and rat), it is obvious that most of these genes code for growth factors expressed during embryonic life either in the fetal membranes, the placenta or in the embryo proper.

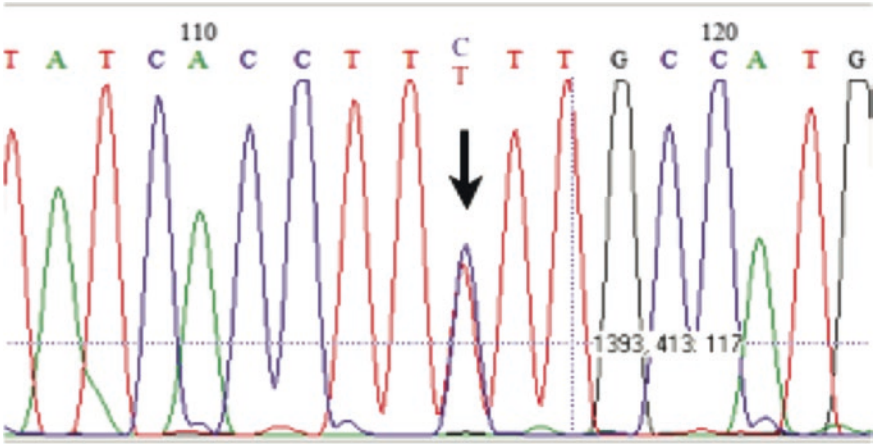


Fig. 6.6 *Molecular identification of imprinted genes using SNPs in the cDNAs*. One can easily check if the two alleles at a given locus are co-expressed in embryonic or adult tissues by analyzing the SNP pattern of the transcribed RNAs. The figure represents part of the sequence of the transcripts of the gene encoding β-hemoglobin (HBB) in the bone marrow cells of F1 mice heterozygous for a single, untranslated nucleotide polymorphism (a silent mutation) in exon 2 of the gene. The figure shows that both alleles are transcribed, since one can recognize the profile of a C/T SNP (*arrow*) in the sequence of the corresponding cDNA. If the gene encoding β-hemoglobin chain (*Hbb*) was among the genes undergoing genomic imprinting, one would have found a single transcript (either from the C or from the T allele) depending on the direction of the cross. Sequencing the whole transcriptome of interspecifc F1 mice is an efficient way of making the inventory of imprinted genes in a given species or in a given tissue

6.3.3 What are the Molecular Mechanisms that Control Genomic Imprinting?

6.3.3.1 DNA Methylation Modifies Transcriptional Activity

Understanding the biological mechanisms involved in the establishment and maintenance of genomic imprinting has motivated a large number of experiments carried out mainly in the mouse and using the most sophisticated techniques. The results obtained have much clarified the situation, even if some aspects require a closer look. We will summarize the state of knowledge as it stands now. However, before doing this, it is important to note that the molecular mechanisms in question had to comply a priori with some basic constraints. First, imprinting may interact with the transcription process but in no way may it alter the DNA sequence of the imprinted regions. Imprinting, as we discussed, is strictly epigenetic, which means that the information in the DNA sequence is not altered. Second, the imprinted regions must be transmitted unchanged to the daughter cells during the development of the embryo and in the adult to ensure the continuation of imprinting, at least for some time, in the different cell lineages. Third, the

epigenetic alteration(s) must be initiated in the paternally or maternally inherited chromosomes independently, and at a time when they are not in the same nucleus; that is, during gametogenesis or immediately after fertilization, before the fusion of the pronuclei. Finally, the parental imprint must be erasable (or reversible) in order to be set differently when the allele goes into a gamete of the opposite sex (Ferguson-Smith 2011).

One of the imprinted regions that has been the most extensively studied is, again, the one that maps to the distal part of mouse chromosome 7. This region, in fact, contains two contiguous clusters: one with four genes (cen–… *H19–Igf2–Igf2as–Ins2*), encompassing around 1 Mb, and another one, more distal, harboring around 15 genes (around the *Kcnq1* locus). Both clusters have a homolog in human and rat, and the genes in question are equally imprinted in these two species.

H19 encodes a 2.3-kb ncRNA that is highly preserved across mammalian species, indicating that it presumably has an important function. Embryos heterozygous for a maternally inherited knockout allele or homozygous for the *H19* knockout allele exhibit increased placenta and body weight (Gabory et al. 2009).

Igf2 encodes a hormone that has similarity with insulin and is probably a major fetal growth factor. Mice heterozygous for an *Igf2* knockout allele ($Igf2^{-}$), transmitted through the male, exhibit pre- and post-natal growth retardation. In contrast, when the disrupted (null) allele is transmitted maternally, the heterozygous offspring are phenotypically normal. Both *H19* and *Igf2* are widely expressed during embryonic development, and then they are down-regulated in most adult tissues.

Shortly after the characterization of the *H19–Igf2* cluster and its complete sequencing, it was demonstrated that imprinting of these two genes is concomitant with the methylation of an *imprinting control region* (ICR) or *differentially methylated region* (DMR), which is 2 kb long and inserted between the two genes. Proper imprinting of *H19* and *Igf2* requires the ICR integrity because, when this region is altered or deleted by genetic engineering, imprinting is abolished. Similarly, proper imprinting of the *H19–Igf2* cluster requires that the ICR be methylated on the paternal allele and unmethylated on the maternal allele.

As already discussed concerning the mechanisms at work in the case of X inactivation, DNA methylation is a biochemical process that consists of the addition of a methyl (CH_3) group at the C-5 position of cytosine, at specific sites known as 5′-C_pG-3′ dinucleotides or C_pG islands. When methylation occurs in the 5′ regulatory regions of many genes, this generally results in transcriptional silencing of these genes. Experiments performed in the early 1990s demonstrated that DNA methylation is probably a crucial step in determining imprinting in mammals, since deficiency in DNA methyltransferase activity (for example, as a consequence of a targeted null mutation at the *Dnmt1* gene) impedes normal imprinting and the homozygote mutant embryos die around day E9.5 (Li et al. 1993). Methylation is stable and can be inherited through mitotic cell division in the differentiated tissues. Methylation alters the spatial conformation of the DNA, making it more compact and accordingly less accessible to DNA-binding proteins, but it does not

alter the sequence proper. Methylation is also reversible and accordingly complies with the constraints mentioned above.[10]

The *H19–Igf2* imprinted region of mouse chromosome 7 and its human homolog on chromosome 11p15.5 have been extensively studied with the aim of elucidating the mechanisms at work for imprinting establishment and maintenance in mammals. Most of the results gathered in the mouse have been cross-validated in humans, and vice versa. As mentioned, these results have revealed the existence of ICRs and DMRs, as regulatory elements for imprinting of the gene cluster, and have underlined the role of methylation of the C_pG islands as previously observed in plants. Methylation of these regions results in silencing or activation of the cluster, depending on the initial status of the genes concerned (Ferguson-Smith et al. 1993; Constância et al. 1998; Reik et al. 2001; Reik and Walter 2001).

The DMRs are the main signature of imprinted genes. Some are called primary or germline DMRs (such as the *H19–Igf2* ICR or the *Igf2r* ICE), because they acquire their differentially methylated status in the germline, and others are called secondary or somatic DMRs and acquire their methylation after fertilization. In the case of the *H19–Igf2* locus, the insulator protein, called CTCF, binds only to the unmethylated ICR and produces a boundary. This results in the interaction of downstream enhancers with the *H19* promoter but not with the *Igf2* promoter on the maternal allele. This was defined as the *enhancer competition model* and explains the monoallelic expression of these genes.

6.3.3.2 Other Mechanisms Involved in the Control of Imprinting

The Role of ncRNAs

Analysis of several imprinted regions also revealed that some specific ncRNAs are probably essential intermediate molecules for the establishment (and maintenance) of imprinting. This assumption was validated by observations made on the imprinted *Igf2r* cluster on mouse chromosome 17. In this cluster, the ICE (imprinting control element) acts as a promoter for a long ncRNA named *Airn* (for antisense of *Igf2r* RNA non-coding) from the unmethylated paternal allele. When *Airn* expression is abolished, the *Igf2r* imprint is removed, suggesting a mechanism of transcription interference (Latos et al. 2012). This mechanism, however, does not exist in humans where *Igf2r* is not imprinted.

[10] Several assays have been designed to assess the methylation status of the genomic DNA. One of the most popular consists of the initial treatment of DNA with sodium bisulfite, which converts cytosine residues into uracil (U) or thymidine (T), but leaves 5-methylcytosine residues unaffected. Once treated with bisulfite the DNA can then be directly sequenced or digested with restriction enzymes (like *Bst*UI), which only cleave sites that were originally methylated (CGCG) but not those that were originally unmethylated (TGTG). Combined bisulfite restriction analysis (or COBRA) is a widespread technique allowing quantification of DNA methylation. It has been extensively used in cancer research and epigenetics studies.

A recent report indicated that within each cluster all imprinted genes show concordant parent-of-origin-specific gene expression except for the ncRNAs that show expression from the opposite parental allele. Such strict reciprocal parent-specific expression seen between mRNAs and imprinted macro ncRNAs strongly indicates that ncRNAs regulate imprinting in such clusters (Saxena and Carninci 2011). This has also been shown for the *Kcnq1* locus, in which the *Kcnq1ot1* long ncRNA is required to maintain DNA methylation and transcriptional gene silencing of the adjacent imprinted genes (Mohammad et al. 2012).

The Role of Histones

Histone modifications have also been considered as an important mechanism in establishing the imprint either directly or indirectly, and in many cases the alleles that display DNA methylation also carry histone marks associated with inactivity. Many points still remain to be clarified concerning the mechanisms of establishment and maintenance of imprinting in mammals (Chen and Dent 2014).

6.3.3.3 Marks of Imprinting are (in General) Cleared Between Generations and Reset During Gametogenesis

The sex-specific marks on DNA, which result in (or lead to) genomic imprinting, and consequently to functional haploidy of the non-imprinted alleles, persist in general from conception throughout all embryonic stages and up to the adult state in most somatic cells. These marks, however, have to be completely erased at a certain critical period of the life cycle since they are likely to be set differently at each generation.

Experiments and observations have demonstrated that epigenetic marks (histone modifications and DNA methylation) on most of the genome start to become erased in primordial germ cells of both sexes at around day 11.5 of gestation, upon entry of the germ cells into the gonads. Genes then acquire new sex-specific DNA methylation marks during fetal development in males and a little later, during the growing oocyte phase, in the early neonatal period in females. The mechanisms involved during the clearing out of the imprinting marks (active or passive DNA demethylation) have not been completely unraveled (Ferguson-Smith 2011).

More importantly, acquired methylation of the ICRs or DMRs of imprinted genes needs to be preserved during the massive wave of demethylation that occurs in the embryo after fertilization. It is now known that imprinted genes display hexanucleotide motifs that are methylated and recognized by several proteins (such as Zfp57, TRIM 28, or Stella). The complex formed between the hexanucleotide motif and these proteins protects the ICRs from being demethylated at these early stages of development and is a signature of the imprinted genes. These observations reveal that both genetic and epigenetic signals are required to establish and maintain the imprinted status of a gene.

When discussing X-chromosome inactivation we mentioned that the inactive X chromosome could sometimes reactivate in somatic cells, especially when the animals age. The situation is similar and even more common with autosomal imprinting, and cases of tissue-specific variations have been reported in the mouse. For example, in the developing embryos only the paternal allele at the *Igf2* locus is expressed, while the maternal allele is silent. However, in the choroid plexus and leptomeninges the situation is different and both alleles are transcriptionally active (DeChiara et al. 1991). Another example of tissue-specific imprinting is provided by the *Cdh15* gene. The germline DMR of this gene is protected from erasure of methylation during the first steps of embryogenesis but becomes methylated after implantation. This led to the proposal of the existence of both *bona fide* imprinted germline DMRs and *transient* germline DMRs (Proudhon et al. 2012).

Another interesting situation is provided by the viable yellow allele at the *agouti* locus (A^{vy}-Chr 2). This mutation is transmitted as a dominant allele; it is viable when homozygous (unlike the classical yellow allele A^{y}, which is homozygous lethal), but the coat color of affected mice exhibits variation, ranging from pure homogeneous yellow, through mottling with dark patches, to an agouti-like coat (pseudo-agouti) similar to the wild-type allele *A*. Homozygous (A^{vy}/A^{vy}) and heterozygous (A^{vy}/a) mice tend to become obese and diabetic, and the degree of obesity is correlated with the coat color, yellow mice being more affected than agouti ones (Morgan et al. 1999).

The A^{vy} mutation is the result of the insertion of an intra-cisternal A-particle (IAP or retrotransposon) into a non-coding exon 5′ of the agouti gene. Functional analysis revealed that the expression of the mutant allele is controlled by the long terminal repeat (LTR) of the IAP. When the LTR in question is hypomethylated, the A^{vy} allele is transcribed, the coat is yellow, and the mouse is bigger than normal. When the viral LTR is methylated (and accordingly inactivated), the coat is agouti. Variegation of coat color in $A^{vy}/+$ mice (which is sometimes also observed in $A^{y}/+$ mice) is very likely the consequence of some mosaicism at the somatic cell level.

When $A^{vy}/+$ males are mated with *a/a* (black non-agouti) females, there is no significant difference in the proportions of yellow, mottled or pseudo-agouti phenotypes in the progenies, and this occurs independently of the coat color (yellow, mottled or agouti) of the male. The situation is different when the cross is set up the other way, i.e., between an *a/a* (non-agouti) male and $A^{vy}/+$ female. In this case there is some sort of *transgenerational epigenetic inheritance* in the sense that the distribution of phenotypes in the progenies is related to the phenotype of the dam and not of the sire—for example, yellow mothers produce more yellow offspring than agouti mothers. Clearly, it appears that imprinting marks are not erased when transmitted through the female, while they are erased when transmitted through the male. Several laboratories have confirmed these observations and it has been demonstrated that selection of a certain phenotype (for example, the percentage of pseudo-agouti offspring in the progeny) could increase the prevalence

of the trait in successive progenies (Blewitt et al. 2006; Cropley et al. 2012). The behavior of the A^{vy} allele, which is quite uncommon in mouse genetics, may appear anecdotal but similar situations might be common if we consider the abundance of IAP in the mammalian genomes (Morgan et al. 1999).

6.3.4 Genomic Imprinting Across Mammalian Species

To date, the differential expression of alleles according to their parental origin has been reported and documented only in flowering plants (Nowack et al. 2007) and in mammals. In mammals, it seems to be an exclusive characteristic of the eutherians and metatherians[11] (marsupials), while prototherians (for example the platypus, *Ornithorhynchus anatinus*) do not exhibit genomic imprinting. In other words, genomic imprinting seems to correlate with gestation of the embryo inside the uterus and placentation (viviparity) but not with egg laying (oviparity). Genomic imprinting has never been reported in fish, amphibians, reptiles or birds (Dünzinger et al. 2005).

In mammals, the imprinted regions are in general relatively well preserved across the different species and for each of the imprinted regions in the mouse, for example, there is in many instances a homologous region in the rat and in humans—with, however, a few remarkable exceptions. From these phylogenetic observations one may conclude that genomic imprinting probably appeared concomitantly with the viviparous mode of reproduction (i.e., ~180 Myr ago). One may also observe that the more closely related are any two species, the greater are the homologies between the different imprinted regions. However, after careful observation it is sometimes discovered that rare but noticeable differences exist between closely related species, as if the process of genomic imprinting was still in evolution in that class of vertebrates.

As we discussed in a previous chapter, some morphological differences between inter-specific hybrids have been reported which depend upon the way the cross that produced these hybrids was set up. Even in the *Mus* genus, in which so many species have been identified including *Mus m. musculus* and *Mus m. domesticus*, some morphological and anatomical differences have been noted that could be attributed to point differences in terms of genomic imprinting. For example, female mice of the *Mus spretus* species do not (or very rarely) produce viable offspring when crossed with laboratory mouse males, while the reverse is not true. The placental hypertrophy of some of these rare F1 hybrids or backcross offspring has been attributed to an X-linked locus (*Ihpd* for interspecific hybrid placental dysplasia) with several alleles, but could also be interpreted as differential imprinting due to differential X inactivation.

[11] In marsupials, the number of imprinted genes is much lower than in eutherian mammals.

6.3.5 The Origin and Evolution of the Imprinting Mechanisms in Mammals

The existence of genomic imprinting raises a number of issues that can be summarized in the following question: what advantage can justify, for a mammalian embryo, having a number of its genes maintained in a functionally haploid status, while diploidy is generally considered more advantageous with regards to evolution? The answer to this basic question is not yet definitively known, and several hypotheses have been developed over the last decade (Wood and Oakey 2006).

One of the first explanations that came to mind was the consideration that imprinting emerged during evolution as a mechanism to clear the genome of spontaneously occurring mutations with lethal or deleterious effects, for the simple reason that such mutations, when they occur within an imprinted region, are eliminated when the region in question becomes functionally haploid. This hypothesis unfortunately has several weaknesses, and in particular it does not explain why such a clever mechanism appeared so late in evolution and has remained an exclusive privilege of mammals.

A more consistent explanation is that genomic imprinting is a very efficient way of inhibiting parthenogenetic (gynogenetic or androgenetic) development in mammals. Indeed, and as explained above, the development of a normal mouse embryo from two female (or two male) pronuclei (i.e., from only one parent or from two parents of the same sex) is strongly repressed. This is a direct consequence of genomic imprinting at the *H19–Igf2* and *Dlk1–Gtl2* loci, as demonstrated by Japanese scientists who succeeded in producing bi-maternal mice after artificially erasing (i.e., by genetic engineering) the imprinting at these loci (Kono et al. 2004; Kawahara et al. 2007; Kawahara and Kono 2012). Although more likely than the previous one, the hypothesis stating that genomic imprinting exists only to impede parthenogenesis in mammals is not entirely convincing and is definitely not sufficient. In fact, the possibility that parthenogenetic development could occur in mammals cannot, a priori, be regarded as a disadvantage, since that sort of development exists occasionally in some classes of vertebrates as an exceptional and alternative way of reproduction, for example to escape a reproductive dead end. From this point of view, the possibility of the mammals using parthenogenesis for one or two generations would also appear advantageous.

A third hypothesis on the origin of genomic imprinting is that it has no advantages at all and exists only by chance. According to this hypothesis imprinting is a mere artifact, a “red herring” so to speak, which results from the uncontrolled expansion to the neighboring regions of a defense mechanism used by mammals to control or neutralize the possible invasion of their genome by self-replicating parasitic DNAs such as retroviruses or retro-transposons (see Chap. 5). Just like the previous two, this hypothesis has some weaknesses and, in particular, it does not explain why imprinting exists only in mammals—while birds have to compete with so many retroviruses and retro-transposons invading their genomes. In the same way, it does not fit with the fact that imprinting is reversible.

If we summarize the information gathered from the observations made in humans (see below) and those collected from the many experiments that have been performed in the mouse species, we can establish correlations and draw some conclusions about the essence of imprinting in mammals and finally come to more coherent hypotheses. An important one is based on the observation that most—not to say all—of the genes that are imprinted have been found to play a role in the control of embryonic growth and development, in most instances through the development of the placenta. Based on this observation, a widely accepted hypothesis to explain the origin and evolution of genomic imprinting is the "*parental conflict hypothesis*", which is also known as the "*tug-of-war hypothesis*". The hypothesis states that the differences between parental genomes due to imprinting are the result of the divergent interests of each parent (or sex) concerning the evolutionary fitness of their genes (Haig 1997; Sha 2008). Since males can have a virtually unlimited number of offspring, the father's genes gain greater fitness through the vigor of the offspring, eventually at the expense of the mother, and this explains why paternally expressed genes tend to be growth-promoting for the embryo. The mother's interest, on the other hand, is to preserve nutrients and resources for her own use, to get rid of the offspring that are in her uterus as soon as possible, and thus be able to produce another litter as rapidly as possible. This would be in agreement with the observation that maternally expressed genes tend to be growth-limiting. Indeed, unlike other vertebrate embryos, mammals could theoretically stay in utero for an unlimited period of time, surviving at the expense of the mother's nutrients, unless a mechanism regulating gestation length intervenes. Genomic imprinting, indirectly controlling the embryo's growth, appears a good way to limit the duration of gestation. This hypothesis has the great advantage of justifying the existence of imprinting and its existence exclusively in mammals and, for this reason, it has been accepted for a good ten years. Nowadays, however, our understanding of the molecular mechanisms at work in genomic imprinting has revealed some inconsistencies, and the parental conflict hypothesis would probably need to be revisited. Recent observations have suggested co-adaptation between the mother and the conceptus at fetal stages (involving placental exchanges) and at post-natal stages with metabolic and behavior exchanges (Keverne 2013).

6.3.6 The Pathological Aspects Associated with Genomic Imprinting

6.3.6.1 Epigenetics and Human Diseases

The same year (1974a) when Johnson reported his observations concerning the phenotypic differences associated with the parental origin of the hairpin-tail (T^{hp}) mutant allele in the mouse (see above), Lubinsky and colleagues reported a similar parental effect in a family transmitting a syndrome now known as

Beckwith–Wiedemann syndrome (BWS) (Lubinsky et al. 1974). In fact, these two observations independently inaugurated the studies relating to the effect of genomic imprinting on gene expression in the mouse and human species, respectively. Nearly forty years after these publications, a lot has been learned concerning genomic imprinting and its importance in some human pathologies.

Beckwith–Wiedemann syndrome (OMIM 130650) is a rare disorder with an incidence of approximately one in 14,000 childbirths. It is characterized by the association of traits like macroglossia, greater than normal birth weight and size, neonatal hypoglycemia, and some other visceral defects (of the adrenal gland in particular). In most cases the BWS is sporadic, but around 15 % of the cases are familial and in many of these familial cases, mutations or deletions of genes within a region spanning approximately 1 Mb of human chromosome 11p15.5 have been reported (the mouse homologous region is on distal chromosome 7). Imprinting defects of genes in the same region have also been described in a very high proportion of BWS patients having a biallelic (rather than paternal monoallelic) expression of the *IGF2* gene. In these cases, the maternal copy of the gene *IGF2* is transcribed where it is normally inactivated in healthy babies. Finally, some babies affected by the BWS have been found to be the consequence of a paternal uniparental disomy (PatUpDi) of chromosome 11, and in these rare cases the two regions of chromosome 11, having escaped maternal imprinting, are both transcribed. Other patients exhibit loss of imprinting of a gene encoding a long ncRNA transcript, called KCNQ1OT1, which is also known to be imprinted in the mouse.

Another rare human syndrome, Russell–Silver syndrome (RSS-OMIM 180860—one in 70,000 childbirths), has also been found to be associated with an imprinting defect. In a recent survey concerning this disease, 10 % of all the cases were found to be associated with a maternal uniparental disomy (MatUpDi) of chromosome 7. In some other cases, the same 11p15.5 region of human chromosome 11 harboring the *H19* and *IGF2* genes appeared to be involved. The defect in this case is characterized by a suppression of IGF2 growth factor activity that explains the concomitant growth reduction observed in RSS patients. In these cases, where the same 11p15.5 region is concerned, the pathological features of RSS logically appear to be the opposite of those described for BWS (Butler 2009).

Prader–Willi (PWS) and Angelman (AS) syndromes are the two most studied cases of human diseases commonly related to defective genomic imprinting. Unlike BWS and RSS, which are often compatible with an almost normal adult life, PWS and AS are always severe and do not improve with aging. PWS and AS are caused by mutations, deletions, uniparental disomy or by abnormal imprinting of one or several different members of a gene cluster in the q11-q13 region of human chromosome 15 (Horsthemke and Wagstaff 2008).

Prader–Willi syndrome (OMIM 176270) occurs in one in 15,000 individuals and is characterized at a young age by hypotonia, short stature, mental deficiency, behavioral problems, and feeding difficulties. In a second phase, from the age of 3 years, developmental delay and psychomotor retardation are even more obvious but obesity becomes a life-threatening issue requiring strict dietary restrictions.

Angelman syndrome (OMIM 105830) is characterized by severe mental retardation with seizures, ataxia, uncoordinated movements, hypopigmentation, inappropriate hilarity, lack of speech, etc. In the late 1980s geneticists observed that PWS and AS were caused by deletions in bands 15q11-q13, and they reported that the observed phenotypic differences between the two syndromes in fact depended upon the parental origin of the deletion. Deletions occurring on paternal chromosome 15 generally resulted in PWS, while similar deletions occurring on maternal chromosome 15 resulted in AS. For this reason, PWS and AS were, and still are, considered as sister syndromes—which fits rather well with the symptomatology.

Nowadays, the situation has been much clarified, and by and large it is concluded that PWS is a consequence of the lack of the paternal copy of one or a few genes in the 15q11-q13 region, while AS is a consequence of the lack of a functional maternal copy of the *UBE3A* gene encoding ubiquitin protein ligase 3A (Moncla et al. 1999; Horsthemke and Wagstaff 2008).

In addition to the four syndromes described above, which are relatively well documented, a few other human diseases and pathological conditions, including certain forms of cancers, have been described as the very likely consequence of abnormal imprinting because of a clear effect of the parental inheritance. In most instances, however, the situation was reported as complex and difficult to analyze because of the interference of environmental factors and/or epistatic interactions with elements of the genetic background. It is likely that, with the rapid progress in sequencing technology and the development of quantitative analysis of RNA transcription, these diseases or syndromes will be clarified in the near future. This will definitely allow a better understanding of the role of epigenetic regulation in gene expression.

6.3.6.2 Epigenetic Manifestations in Some Animal Crosses

At several points in this book we have mentioned that some interspecific mouse hybrids exhibit a variety of pathological features depending on the direction of the cross. For example, crosses between male mice of the *Mus spretus* species and females of the *Mus m. domesticus* species produce viable hybrids but the sex ratio in the offspring progeny is much biased in favor of the female, and the male F1s are always sterile. This difference is in compliance with the so-called Haldane's rule and has been observed in several other cases of interspecific crosses (for example, between different *Drosophila* species, between *Bos taurus* and *Bison bison*, and between *Chrysolophus pictus* and *Gallus g. domesticus*).[12] In the case of mouse crosses, it has been established that the sterility of hybrids is controlled by a few genes, some of which have been localized on the genetic map. In contrast, the reasons for the shortage of males are still conjectural.

[12] Haldane's rule states "when in the offspring of two different animal races one sex is absent, rare, or sterile, that sex is the heterozygous [heterogametic] sex."

More interesting is the observation that crosses in the other direction (between *Mus m. domesticus* males and *Mus spretus* females) result in stillbirths in most cases, with a marked enlargement of the placenta.[13] A similar situation was reported for crosses between two other species of rodents of the genus *Peromyscus*, with strong parent-of-origin effects involving placental growth. Female *P. maniculatus* crossed with male *P. polionotus* produce neonates smaller than either parental strain, with placentas half the parental size. In contrast, female *P. polionotus* crossed with male *P. maniculatus* produce dysmorphic overgrown embryos whose placentas average up to 2.5 times the mass of the parental strains (Vrana 2007).

Such biases are difficult to explain in terms of Mendelian genetics if we consider that the genetic makeup of the above-mentioned reciprocal F1s are virtually the same, with one allele of each parental species in both cases. However, a possible (and likely?) explanation would be to guess that the parental alleles of some homologous genes are imprinted differently in the two F1s. This would explain all the observed phenotypes.

A similar observation has been made concerning the offspring of crosses made in zoological gardens between two species of the *Panthera* genus: *Panthera leo*, the African lion, and *Panthera tigris*, the Bengal tiger. The *liger*, a hybrid between a male lion and a tigress, is an enormous animal, with a total length reaching 3–3.5 m and a weight of up to 380 kg (~800 lb), while the reciprocal hybrid, the *tigon* (much less common), is slightly undersized compared to its parents. Here again, the explanations for these size differences are still somewhat speculative but, given that the imprinted genes often play a role in issues of hybrid growth, it is tempting to guess that this applies in the case of these two interspecific hybrids (Morison et al. 2001, 2005).

Finally, another interesting case is the *Callipyge* phenotype in sheep (abbr. *CPLG*—from the Greek "beautiful buttocks"). This mutation was first discovered in the USA segregating in a flock in Oklahoma. It causes lambs to develop large and muscular rumps, and for this important economical value it has been extensively studied by animal geneticists (Georges et al. 2013). It has then been demonstrated that the phenotype is fully expressed only in heterozygous individuals who receive the *CLPG* mutant allele from their father. When inherited from the mother, it is not expressed. This situation is known as *polar overdominance* and is another example of phenotypic alteration due to imprinting. The *CLPG* mutation is a single nucleotide substitution in what is probably a long-range control element (LRCE—see Chap. 5) within the *DLK1–GTL2* imprinted domain of several species of mammals. The mutation also exists in humans and in cattle, and has been created by genetic engineering in the mouse. It is a very interesting model for these sorts of phenotypic observations.

[13] Only some exceptional viable offspring have been bred from such a cross.

6.4 Conclusions

Initially discovered in the form of anecdotal observations (coat color of calico cats and the unexpected inheritance of the hairpin-tail mutation), X inactivation and genomic imprinting appear to be two important ways of regulating genomic expression. Diploidy, as we said, was generally considered as advantageous with regards to evolution because, having a backup copy for each and every gene, diploid organisms were more protected against the deleterious effects of mutations. After the discovery of genomic imprinting, this analysis must be seriously reconsidered. Indeed, if a gene mutates, the back-up (normal) copy of this gene may not be available for replacement if it is in an imprinted region and accordingly epigenetically inactivated. What then is the evolutionary advantage of imprinting for mammals? A close association has been established with viviparity, at least with the development of the embryo in utero, but this association by definition does not exist in flowering plants where the imprinting phenomenon has also been described. Nowadays, a theory is emerging suggesting that genomic imprinting might play an important role as a mechanism of reproductive isolation generating diversity. Many of these investigations are conducted in mammals (in particular, laboratory rodents), and it is likely that the evolutionary advantages of genomic imprinting will be established in the relatively near future.

Unraveling the intimate molecular mechanisms at work in the establishment and maintenance of imprinting might be laborious, but it is a very important issue and there is no doubt that, in this matter more than in any other, the mouse will be an invaluable model.

References

Augui S, Nora EP, Heard E (2011) Regulation of X-chromosome inactivation by the X-inactivation centre. Nat Rev Genet 12:429–442

Barlow DP, Stöger R, Herrmann BG, Saito K, Schweifer N (1991) The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature 349:84–87

Barton SC, Surani MAH, Norris ML (1984) Role of paternal and maternal genomes in mouse development. Nature 311:374–376

Beutler E, Yeh M, Fairbanks VF (1962) The normal human female as a mosaic of X-chromosome activity: studies using the gene for G-6-PD deficiency as a marker. Proc Natl Acad Sci USA 48:9–16

Blewitt ME, Vickaryous NK, Paldi A, Koseki H, Whitelaw E (2006) Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. PLoS Genet 2(4):e49

Brown SD (1991) XIST and the mapping of the X chromosome inactivation centre. BioEssays 13:607–612

Butler MG (2009) Genomic imprinting disorders in humans: a mini-review. J Assist Reprod Genet 26:477–486

Cattanach BM (1986) Parental origin effects in mice. J Embryol Exp Morphol 97(Suppl):137–150

Cattanach BM, Kirk M (1985) Differential activity of maternally and paternally derived chromosome regions in mice. Nature 315:496–498

Chen T, Dent SY (2014) Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat Rev Genet 15:93–106
Constância M, Pickard B, Kelsey G, Reik W (1998) Imprinting mechanisms. Genome Res 8:881–900
Cropley JE, Dang TH, Martin DI, Suter CM (2012) The penetrance of an epigenetic trait in mice is progressively yet reversibly increased by selection and environment. Proc Biol Sci B 279:2347–2353
Davidson RG, Nitowsky HM, Childs B (1963) Demonstration of two populations of cells in the human female heterozygous for glucose-6-phosphate dehydrogenase variants. Proc Nat Acad Sci USA 50:481–485
DeChiara TM, Robertson EJ, Efstratiadis A (1991) Parental imprinting of the mouse insulin-like growth factor II gene. Cell 64:849–859
Dünzinger U, Nanda I, Schmid M, Haaf T, Zechner U (2005) Chicken orthologues of mammalian imprinted genes are clustered on macrochromosomes and replicate asynchronously. Trends Genet 21:488–492
Espinós C, Lorenzo JI, Casaña P, Martínez F, Aznar JA (2000) Haemophilia B in a female caused by skewed inactivation of the normal X-chromosome. Haematologica 85:1092–1095
Ferguson-Smith AC (2011) Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet 12:565–575
Ferguson-Smith AC, Sasaki H, Cattanach BM, Surani MA (1993) Parental-origin-specific epigenetic modification of the mouse H19 gene. Nature 362:751–755
Gabory A, Ripoche MA, Le Digarcher A, Watrin F, Ziyyat A, Forné T, Jammes H, Ainscough JF, Surani MA, Journot L, Dandolo L (2009) H19 acts as a trans regulator of the imprinted gene network controlling growth in mice. Development 136:3413–3421
Garrick D, Sharpe JA, Arkell R, Dobbie L, Smith AJH et al (2006) Loss of *Atrx* affects trophoblast development and the pattern of X-inactivation in extraembryonic tissues. PLoS Genet 2(4):e58
Georges M, Charlier C, Cockett N (2013) The callipyge locus: evidence for the trans interaction of reciprocally imprinted genes. Trends Genetics (in press)
Haig D (1997) Parental antagonism, relatedness asymmetries, and genomic imprinting. Proc Roy Soc Lond Ser B-Biol Sci 264:1657–1662
Horsthemke B, Wagstaff J (2008) Mechanisms of imprinting of the Prader-Willi/Angelman region. Am J Med Genet A 146:2041–2052
Johnson DR (1974a) Further observations on the haipin-tail (T^{hp}) mutation in the mouse. Genet Res 24:207–213
Johnson DR (1974b) Hairpin-tail: a case of post-reductional gene action in the mouse egg. Genetics 76:795–805
Kawahara M, Kono T (2012) Roles of genes regulated by two paternally methylated imprinted regions on chromosomes 7 and 12 in mouse ontogeny. J Reprod Dev 58:175–179
Kawahara M, Wu Q, Takahashi N, Morita S, Yamada K, Ito M, Ferguson-Smith AC, Kono T (2007) High-frequency generation of viable mice from engineered bi-maternal embryos. Nat Biotechnol 25:1045–1050
Kelly WG, Schaner CE, Demburg AF, Lee MH, Kim SK, Villeneuve AM, Reinke V (2002) X-chromosome silencing in the germline of *C. elegans*. Development 129:479–492
Keverne EB (2013) Importance of the matriline for genomic imprinting, brain development and behaviour. Philos Trans R Soc Lond B Biol Sci 368:20110327. doi:10.1098/rstb.2011.0327
Kono T, Obata Y, Wu Q, Niwa K, Ono Y, Yamamoto Y, Park ES, Seo JS, Ogawa H (2004) Birth of parthenogenetic mice that can develop to adulthood. Nature 428:860–864
Larschan E, Bishop EP, Kharchenko PV, Core LJ, Lis JT, Park PJ, Kuroda MI (2011) X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. Nature 471:115–118
Latos PA, Pauler FM, Koerner MV, Şenergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE, Aumayr K, Pasierbek P, Barlow DP (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science 338:1469–1472

Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. Nature 366:362–365
Lubinsky M, Herrmann J, Kosseff A, Opitz JM (1974) Autosomal-dominant sex-dependent transmission of the Beckwith-Wiedemann syndrome. Lancet 1:932
Lyon MF (1961) Gene action in the X-chromosome of the mouse (*Mus musculus L.*). Nature 190:372–373
Lyon MF (2002) A Personal History of the Mouse Genome. Annu Rev Genomics Hum Genet 3:1–16
Marahrens Y, Panning B, Dausman J, Strauss W, Jaenisch R (1997) Xist-deficient mice are defective in dosage compensation but not spermatogenesis. Genes Dev 11:156–166
McGrath J, Solter D (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell 37:179–183
Mohammad F, Pandey GK, Mondal T, Enroth S, Redrup L, Gyllensten U, Kanduri C (2012) Long noncoding RNA-mediated maintenance of DNA methylation and transcriptional gene silencing. Development 139:2792–2803
Moncla A, Malzac P, Livet MO, Voelckel MA, Mancini J, Delaroziere JC, Philip N, Mattei JF (1999) Angelman syndrome resulting from UBE3A mutations in 14 patients from eight families: clinical manifestations and genetic counseling. J Med Genet 36:554–560
Morey C, Avner P (2011) The demoiselle of X-inactivation: 50 years old and as trendy and mesmerising as ever. PLoS Genet 7:e1002212
Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23:314–318
Morison IM, Paton CJ, Cleverley SD (2001) The imprinted gene and parent-of-origin effect database. Nucleic Acids Res 29:275–276
Morison IM, Ramsay JP, Spencer HG (2005) A census of mammalian imprinting. Trends Genet 21:457–465
Nowack MK, Shirzadi R, Dissmeyer N, Dolf A, Endl E, Grini PE, Schnittger A (2007) By-passing genomic imprinting allows seed development. Nature 447:312–315
Okamoto I, Otte AP, Allis CD, Reinberg D, Heard E (2004) Epigenetic dynamics of imprinted X inactivation during early mouse development. Science 303:644–649
Patrat C, Okamoto I, Diabangouaya P, Vialon V, Le Baccon P, Chow J, Heard E (2009) Dynamic changes in paternal X-chromosome activity during imprinted X-chromosome inactivation in mice. Proc Natl Acad Sci USA 106:5198–5203
Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. Nature 379:131–137
Pollex T, Heard E (2012) Recent advances in X-chromosome inactivation research. Curr Opin Cell Biol. http://dx.doi.org/10.1016/j.ceb.2012.10.007
Proudhon C, Duffié R, Ajjan S, Cowley M, Iranzo J, Carbajosa G, Saadeh H, Holland ML, Oakey RJ, Rakyan VK, Schulz R, Bourc'his D (2012) Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. Mol Cell 47:909–920
Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. Science 293:1089–1093
Reik W, Walter J (2001) Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote. Nat Genet 27:255–256
Saxena A, Carninci P (2011) Whole transcriptome analysis: what are we still missing? Wiley Interdiscip Rev Syst Biol Med 3:527–543
Sha K (2008) A mechanistic view of genomic imprinting. Annu Rev Genomics Hum Genet 9:197–216
Simmler MC, Cattanach BM, Rasberry C, Rougeulle C, Avner P (1993) Mapping the murine Xce locus with $(CA)_n$ repeats. Mamm Genome 4:523–530
Surani MAH, Barton SC, Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature 308:548–550

Thorvaldsen JL, Krapp C, Willard HF, Bartolomei MS (2012) Nonrandom X chromosome inactivation is influenced by multiple regions on the murine x chromosome. Genetics 192:1095–1107

Trent S, Dennehy A, Richardson H, Ojarikre OA, Burgoyne PS, Humby T, Davies W (2011) Steroid sulfatase-deficient mice exhibit endophenotypes relevant to attention deficit hyperactivity disorder. Psychoneuroendocrinology 37:221–229

Vrana PB (2007) Genomic imprinting as a mechanism of reproductive isolation in mammals. J Mammal 88:5–23

Weidman JR, Dolinoy DC, Maloney KA, Cheng JF, Jirtle RL (2006) Imprinting of opossum *Igf2r* in the absence of differential methylation and Air. Epigenetics 1:49–54

Winking H, Silver LM (1984) Characterization of a recombinant mouse T haplotype that expresses a dominant lethal maternal effect. Genetics 108:1013–1020

Wood AJ, Oakey RJ (2006) Genomic imprinting in mammals: emerging themes and established theories. PLoS Genet 2:e147

Wutz A, Theussl HC, Dausman J, Jaenisch R, Barlow DP, Wagner EF (2001) Non-imprinted *Igf2r* expression decreases growth and rescues the *Tme* mutation in mice. Development 128:1881–1887

Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min JH, Jin P, Ren B, He C (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149:1368–1380

Chapter 7
Mutations and Experimental Mutagenesis

7.1 The Importance of Mutations

The word mutation was coined in 1901 by Hugo De Vries to describe "*sudden, spontaneous and drastic alterations in the hereditary material of Oenothera"*, the evening primrose.[1] Mutations occur in the genome of all living organisms and vary in importance, ranging from single base-pair changes to extensive chromosomal rearrangements. They can occur either in somatic or germ cells, at all stages of development, and are transmitted to daughter cells except when they cause death or a severe selective disadvantage.

When mutations occur in somatic cells with high mitotic activity, such as cells of the bone marrow, intestinal mucosa, lung or skin, or when the mutations in question interfere with the mechanisms that regulate the cell cycle or differentiation, then the affected cells may become carcinomatous. When mutations occur in the cells contributing to the germ line they may be transmitted to the next generation and, in this case, a proportion of the offspring will be heterozygous for a new mutant allele. This category of mutations is precisely the one we will focus on in this chapter.

Germinal mutations, by definition, generate new alleles that enter the gene pool of the species and contribute to an increase in polymorphism. Most of these new alleles have no effects or effects that do not influence the fitness of the affected individuals, and for this reason they are called "*neutral mutations"*. A small proportion of these new mutations may result in a better adaptation of the animals to their environment.[2] Finally, some mutations have deleterious effects, frequently leading to pathological conditions. In this case, and if we consider that almost all genes in the mouse genome have an equivalent in the human genome, it is obvious that many among the new mutant alleles found in the mouse species represent potentially interesting models of human genetic diseases.

---

[1] Hugo de Vries also used the word "*sport*" to define the same sort of sudden genetic changes.

[2] Resistance to the rodenticide warfarin is a good example of the mutations that occurred in wild populations, generating a selective advantage.

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_7

Mutations can affect all genomic regions, with a wide range of consequences at the phenotypic level. They are either dominant, semi-dominant (heterozygotes have a less severe phenotype than homozygous mutants), co-dominant (both alleles are equally expressed) or recessive. Detailed study of the phenotype of these new mutant alleles is part of the process of genome annotation, and is of great importance for the characterization of gene function(s).

The occurrence of spontaneous mutations in mammalian genomes results from errors occurring either during meiosis or in the process of DNA replication which are not mended by the cellular (DNA) repair mechanisms. These repair mechanisms are very sophisticated, with specific enzymes constantly checking the integrity of cellular DNA during and after replication, but the system is sometimes defective or saturated and fails. Taking this into account, one understands that there is no way to prevent mutations from occurring, and that the spontaneous mutation rate is a basic parameter that each species must cope with. In addition, many agents such as radiation, some chemicals, and some viruses and transposons can increase the rate of mutations well above the spontaneous rate. Some of these agents, as we will discuss in this chapter, have been used over the last fifty years for performing *experimental mutagenesis*.

Experimental mutagenesis can be *"phenotype-driven"*, when unknown genes are identified based on the phenotypic changes associated with at least one of the mutant alleles. In this case, the structure of the gene affected by the mutation is elucidated afterwards, by positional cloning, depending on the potential interest of the mutant allele. Experimental mutagenesis can also be "*genotype- or gene-driven"*, whereby mutations are massively induced and then sought only in pre-selected genes or DNA regions of unknown function, for example for the purpose of genome annotation. As we will see, experimental mutagenesis is relatively simple to achieve, but its efficiency depends upon the mutagenic treatment as well as on the protocols used for the characterization of the mutant phenotypes.

In this chapter, we will describe in some detail the different types of mutations that can affect a mammalian genome and their consequences. We will then discuss the different protocols that can be used for the induction of mutations in the mouse germline, with special emphasis on chemical mutagenesis, which is highly efficient and accordingly has become widespread.

7.2 The Different Types of Mutations

When considered at the DNA level, mutations are generally classified into two categories:

- *chromosomal mutations*, which are detectable by the observation of morphological changes at the karyotype level, and
- *point mutations*, when no alteration in chromosome integrity is detectable.

This classification into chromosomal mutations and point or gene mutations dates back to a time when the microscope was the only tool available to visualize changes

in the hereditary material. Since then, the notion of point mutation has changed and now covers a group of structurally defined changes occurring in the DNA. We will describe these changes, from the simplest to the more complex, and in so doing we will realize that the classification mentioned above, in fact, is not really stringent. However, it is convenient from a didactic point of view and thus we will adopt it.[3]

The geneticist H.J. Muller,[4] who did pioneering research on experimental mutagenesis in *Drosophila* flies using X-rays, proposed a classification of the mutations into five categories based on the effect of the genetic change on gene activity. The first category, the *amorphic* mutations, consisted of those mutations that completely abolish the activity of the gene and were equivalent to *null* or *loss of function alleles*. *Hypomorphic* mutations were associated with reduced activity compared to the wild-type allele, while *hypermorphic* mutations were the opposite, with an increased activity. *Neomorphic* mutations were the group of mutations exhibiting a new function, and *antimorphic* alleles were mutations with dominant negative effects.

7.2.1 Mutations Resulting from Base-Pair Substitutions in the Coding Sequences

The information gathered from recent efforts of systematic sequencing of the genome of various mouse strains have revealed that nucleotide substitutions are the most frequent type of mutations. We will then take a simple example of this type of mutation and discuss its consequences. This example will be the DNA codon 5′-TGT-3′, which is transcribed as UGU and encodes the amino acid cysteine (Cys, or C when using the one-letter code) and, for the time being, we will focus on the nucleotide at the third position of this codon (T)[5] (Table 7.1).

The first substitution is when the thymine (T) in the DNA strand is replaced by a cytosine (C) (this substitution of a pyrimidine for another pyrimidine is called a *transition*). In this case, the resulting mRNA codon becomes UGC but, like UGU, it still encodes the same Cys residue. This type of mutation has no effect on the protein sequence, and for this reason, it is said to be silent or *synonymous*. In this case, the base substitution is detected only when it suppresses or creates a restriction site or, when comparing sequences, as a single nucleotide polymorphism (SNP).

---

[3] Chromosomal rearrangements such as translocations of all types, transpositions, deletions, duplications, inversions, etc. have been discussed in Chap. 3 (*Cytogenetics*) and 5 (*The Mouse Genome*).

[4] Hermann J. Muller was awarded the Nobel Prize in Physiology or Medicine for the "*discovery that mutations can be induced by X-rays*".

[5] Here we write the codon sequences as they are read on the sense (5′ to 3′) strand of DNA. In these conditions, they read the same as the RNA codons (with the exception that T is replaced by U in mRNAs). However, it must be kept in mind that the mRNA transcripts are synthesized using the antisense strand of DNA (3′ to 5′) as a template.

Table 7.1 Point mutations in the coding sequences

(a)

3'-	CAC	GTG	GAC	ACA	GGA	CTC	CTC	TTC -5'
5'-	GTG	CAC	CTG	TGT	CCT	GAG	GAG	AAG -3'
	↓	↓	↓	↓	↓	↓	↓	↓
5'-	GUG	CAC	CUG	UGU	CCU	GAG	GAG	AAG -3'
	Val	His	Leu	Cys	Pro	Glu	Glu	Lys

(b)

3'-	CAC	GTG	GAC	ACG	GGA	CTC	CTC	TTC -5'
5'-	GTG	CAC	CTG	TGC	CCT	GAG	GAG	AAG -3'
	↓	↓	↓	↓	↓	↓	↓	↓
5'-	GUG	CAC	CUG	UGC	CCU	GAG	GAG	AAG -3'
	Val	His	Leu	Cys	Pro	Glu	Glu	Lys

(c)

3'-	CAC	GTG	GAC	ACT	GGA	CTC	CTC	TTC -5'
5'-	GTG	CAC	CTG	TGA	CCT	GAG	GAG	AAG -3'
	↓	↓	↓	↓	↓	↓	↓	↓
5'-	GUG	CAC	CUG	UGA	CCU	GAG	GAG	AAG -3'
	Val	His	Leu	STOP	Pro	Glu	Glu	Lys

(d)

3'-	CAC	GTG	GAC	ACC	GGA	CTC	CTC	TTC -5'
5'-	GTG	CAC	CTG	TGG	CCT	GAG	GAG	AAG -3'
	↓	↓	↓	↓	↓	↓	↓	↓
5'-	GUG	CAC	CUG	UGG	CCU	GAG	GAG	AAG -3'
	Val	His	Leu	Trp	Pro	Glu	Glu	Lys

This table represents the two strands of a short coding DNA sequence, with three examples of nucleotide substitution on the third position of the ACA/TGT codon, and their consequences at the protein level. **a** The original (non-mutated) DNA strand. **b** The TGT–TGC substitution (a transition) has no consequence at the protein level due to the degeneracy of the genetic code. In this case, the same cysteine residue is incorporated into the protein—it is a synonymous mutation. **c** The replacement of a TGT by a TGA (a transversion), on the contrary, leads to the termination of the translation process; this is a nonsense mutation. **d** Finally, the replacement of a TGT by a TGG codon (another transversion) results in the incorporation of a different amino acid with a wide variety of possible consequences; this is a missense mutation

Synonymous mutations are common findings (~23 %) in the sequence databases of the different mouse inbred strains, especially those recently derived from wild specimens (Beier 2000; Sakuraba et al. 2005; Frazer et al. 2007; Keane et al. 2011; Yang et al. 2011; Arnold et al. 2012), and this observed frequency is in keeping with theoretical computations. Indeed, if we consider that each of the 61 sense codons

(64 – 3) can mutate to one of nine different codons after the substitution of one or the other of the three nucleotides, we can calculate that out of these 549 (61 × 9) possible mutations around 25 % are synonymous while most others (75 %) are not (Graur 2003). If we take a closer look at the distribution of all these mutations we may notice that the synonymous mutations are much more frequent when the substitutions occur on the third nucleobase of the codon (70 %) than when they affect one of the other two positions. This is, of course, because the code is degenerated.

Synonymous mutations occur constantly and regularly, even if at a low rate. They are also relatively stable and have virtually no impact on the phenotype. For these reasons they represent an interesting class of polymorphism for evolutionists and can be considered as a molecular clock useful, for example, for assessing the time of divergence between any two species or strains (Gilman 1972).[6]

These synonymous SNPs, when considered with the other flanking SNPs of the same type on the same chromosome, can also be used for identifying the phylogenetic origin of the chromosome (or haplotype) in question. We will come back to this point when discussing the inheritance of complex or quantitative traits (Keane et al. 2011) (Chap. 10).

An interesting observation is that, in mammals, some synonymous codons are found more frequently than others, even when the codons in question encode the same amino acid. For example, the 5'-AGA-3' and 5'-AGG-3' DNA codons both encode the amino acid arginine (R), but AGA is six times more frequent than AGG in the transcripts. A similar observation can be made with the codons 5'-ACA-3' and 5'-ACG-3', which both encode the amino acid threonine (T), but ACA is five times more frequent than ACG in the transcripts. The reason(s) for such a bias in codon use is (are) not yet elucidated: they may be related to the fact that the mutation rate is not the same for the four different nucleotides (discussed later); alternatively, the bias in the codon usage may be related to the fact that the synonymous codons are not equivalent in terms of efficiency at the translational level; some of the codons have a selective advantage over the others.

Let us now assume that the third nucleotide of the same 5'-TGT-3' codon, the T, is replaced by an adenine (A)—this change is designated a *transversion* (i.e., the substitution of a pyrimidine for a purine). This mutation results in the incorporation of the UGA codon instead of UGU in the mRNA transcript, but this is the signal for the termination of polypeptide synthesis, or stop codon. The resulting mutations are called *nonsense* mutations, generating null or non-functional alleles. Analysis of the sequencing data from positional cloning (in human and mouse) of mutant alleles with a deleterious effect reveals that mutations of this type represent around 4–5 % of the overall point mutations found in the coding sequences.

The functional consequences of nonsense mutations depend on the type of protein encoded by the gene and the potential existence of other genes capable of achieving the same or similar function(s). If the protein has an important function in cellular metabolism and if the gene is present as a single copy, the mutation

[6] The average spontaneous mutation rate at the DNA level has been estimated to be 2.2×10^{-9} per nucleotide per year in the human species (Kumar and Subramanian 2002).

generally leads to cell and/or embryonic death when in the homozygous state (recessive lethal). If, however, the encoded protein is not essential or if it is expressed only in a limited number of cells—for example, only the cells that are involved in the synthesis of melanin pigment (melanocytes)—then only the hair coat and retina of the animal are affected by the mutation, resulting in albinism (the consequence of a null allele of the tyrosinase-encoding gene *Tyr*-Chr 7). All intermediates between these two extreme cases are possible. Typically, inactive alleles resulting from a stop codon have no phenotypic expression when heterozygous, except in the case of haplo-insufficiency or parental imprinting of the normal allele (see Chap. 6).

mRNAs with a premature stop codon are in general rapidly degraded by specific exonucleases.[7] However, in some cases where the stop codon occurs close to the 3' end of the gene (in the last exon, for example), the transcript often escapes mRNA decay and the abnormal (truncated) protein may have a dominant negative effect of variable intensity.

The reciprocal mutations, where one of the three stop codons 5'-TAA-3', 5'-TGA-3', and 5'-TAG-3' reverts to a non stop-codon, are called *read-through mutations*. These mutations are exceptional and only a very small number have been reported (Noveroske et al. 2000). This is understandable if we consider the relatively small target the three stop codons represent (9 bp altogether) compared to the rest of the exonic sequences.

The last substitution we must consider is when the third base of the codon 5'-TGT-3' for thymine (T) is replaced by a guanine (G); this change is another transversion. This substitution changes the mRNA codon UGU to UGG, and a different amino acid (Trp—tryptophan or W) is inserted into the polypeptide chain instead of the original cysteine. These mutations are called *non-synonymous* or *missense,* and their effects are almost unpredictable because they depend upon the site where the substitution occurred and the type of amino acid replacement. This sort of mutation is by far the most common type found in sequencing data from positional cloning of mutant alleles with a deleterious effect. In some cases, the change has extremely limited effects and only some biophysical characteristics of the protein, such as, for example, its electrical charge, are altered. In the case of altered electrical charge, the proteins are designated electrophoretic variants; they are easily identified by electrophoresis in a non-denaturing gel, but the function of the protein remains generally unchanged (see Chap. 4).

The β-chain of mouse hemoglobin (HBB, encoded by the *Hbb* gene on Chr 7) has been extensively studied in wild mice because it represents an interesting system for evaluating the functional divergence of duplicated genes during evolution. In these studies, it has been observed that amino acid changes in the β-globin chain are very common among the different species that are close relatives of the laboratory mice in the genus *Mus*, but all these "mutant" molecules (called isoforms) are perfectly functional (Runck et al. 2009).

Another example of a non-synonymous mutation is worth mentioning: the $Tyr^{c\text{-}h}$ or Himalayan allele at the *Tyr* locus in the mouse. This spontaneous mutation is common

[7] This is referred as nonsense-mediated mRNA decay.

in mammals and an orthologous mutant allele also exists in the rat, the Siamese breed of cats, the rabbit, and several other mammalian species. In the mouse, the mutation was found to be the consequence of an A → G transition at nucleotide 1,259 of the *Tyr* gene, which results in an amino acid change at position 420 from histidine to arginine (His420Arg—a structurally important change). Because of this mutation, melanin synthesis in $Tyr^{c\text{-}h}/Tyr^{c\text{-}h}$ homozygous mice becomes temperature-sensitive; the pigment is synthesized normally in the fur at around 20 °C but not at ~30 °C. As a result, the mice have a different fur color at their extremities (the tip of their nose, tail, limbs, and ear are normally pigmented because the temperature is lower at these parts of the body, while the rest of the mouse is not or weakly pigmented). The Himalayan allele, which is of ancient origin, has been relatively easy to detect and propagate because it made the mice quite eye-catching without altering their health. However, if such mutations occur in genes encoding proteins with an important role in homeostasis of the organism, the consequences, although unpredictable, might be severe.

So far we have only considered the mutations that are the consequences of substitutions occurring at the third position of the 5′-TGT-3′ DNA codon. This codon was selected as an example because it is one of the rare types that can produce the three classes of mutations (synonymous, nonsense, and missense) with a single base-pair replacement at the same position. However, as we already mentioned and because of the degeneracy of the genetic code, mutations at the first and second nucleotides of mRNA codons are generally more deleterious in terms of consequences than mutations at the third position. Using the same permutation as explained above, we can calculate, for example, that substitutions at the first or second position would generate a missense mutation in 91 and 96 % of the cases, respectively.[8] Even if this theoretical computation must be corrected, taking into account that the nucleotides are not represented in equal proportions in the mouse DNA, and accordingly that all 64 codons are not equally frequent, this percentage of non-synonymous mutations is very close to the data actually collected after positional cloning of hundreds of mutations and analysis of mouse genome sequences.[9]

Although predictions concerning the possible deleterious effects associated with missense mutations are difficult and always depend on the genomic context, a number of observations that have accumulated over time provide some indications. For example, it has been observed repeatedly that non-synonymous mutations replacing an aliphatic amino acid with an aromatic one (for example TCG → TGG) have deleterious consequences in most cases. The same applies to the mutations replacing one of the two amino acids containing a sulphur (S) atom (Cys or Met) by another amino acid not containing the S atom. Most amino acid substitutions occurring in the highly conserved domains of proteins almost always have deleterious consequences. Finally, missense mutations leading to an important structural change at the C-terminus often have severe effects by hampering the

[8] Four substitutions of the first nucleobase result in a synonymous codon (lysine or arginine codons). No substitution of the second nucleobase leads to a synonymous codon.

[9] In mouse nuclear DNA, the G + C content is 41.70 %, indicating that codons making use of these two nucleotides are under-represented in this species.

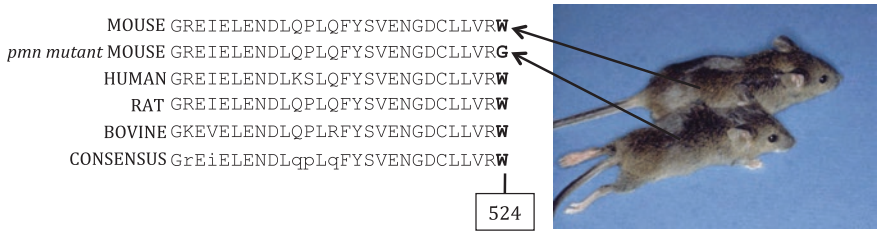


Fig. 7.1 *Missense mutations*. The severe mouse neurological syndrome called progressive motor neuronopathy is the consequence of a missense mutation (*Tbce*^{pmn}-Chr 13) affecting the gene encoding the tubulin-specific chaperone E protein (TBCE). This missense mutation leads to the replacement of the very last amino acid of the protein, a tryptophan residue at position 524, by a glycine (in *short*: Trp524Gly). This change, which is unique to the mutant mouse and is not found in any other species, has consequences for the stability of the protein, and this probably explains the relatively late onset of the pathology (adapted from Martin et al. 2002)

correct folding of the protein, as is the case for progressive motor neuronopathy (*Tbce*^{pmn}) (Fig. 7.1).

Accumulation of new data of this kind contributes to the enrichment of databases, and all of these findings are important for a better understanding of the molecular mechanisms leading to genetic diseases. In this matter, it must be kept in mind that the information gathered from observations made in the mouse are universal and accordingly apply to all mammalian species. In human, around 56 % of the mutations resulting in a pathology are point mutations of the nonsense or missense types. Analysis of a large number of nucleotide substitutions associated with disorders shows that the most common substitutions are T to C, C to T, A to G, and G to A (Krawczak et al. 1998). In humans, the most common type of single nucleotide substitution is the C_pG dinucleotide that mutates to T_pG at a frequency which is about five times higher than mutations in all other dinucleotides (Youssoufian et al. 1988; Antonarakis et al. 1995; Krawczak et al. 1998). There is no reason to think that this frequency might be different in the mouse.

7.2.2 Base-Pair Substitutions in the Non-coding Regions

Base-pair substitutions in non-coding regions of the genome are innumerable, and the data gathered from mouse, rat, and human sequencing efforts provide many examples of such substitutions that, in most instances, have been recorded as mere SNPs with no detectable phenotypes. Exceptions are when the changes occur in splicing sites or in regulatory regions. These two kinds of mutations represent, respectively, 9.3 % and 1.9 % of the mutations associated with a pathological syndrome in humans, and it is likely that the proportion is similar in the mouse.

Mutations that interfere with the splicing process result in *exon skipping* or in the reciprocal defect known as *intron retention*. In some other instances, a cryptic splicing site is activated after a single base-pair substitution, and this results in the incorporation of a DNA segment of intronic origin into the transcript and possibly into the encoded protein (Figs. 7.2 and 7.3).

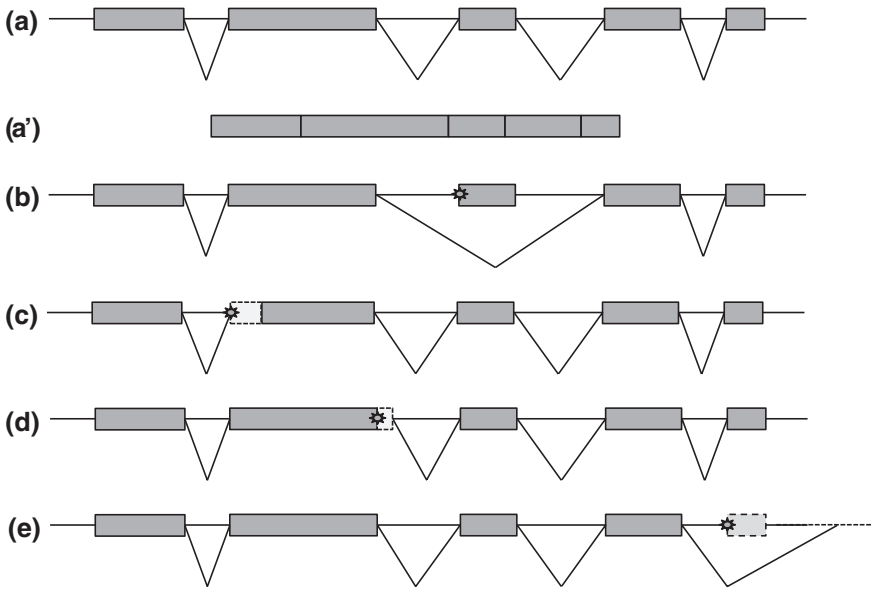


Fig. 7.2 *Examples of splicing defects generated by nucleotide substitutions.* **a** Schematic representation of a normal gene. Exons are shown as grey boxes and introns as lines between exons. a′ represents the mature mRNA transcript after splicing of all introns. **b** A nucleotide substitution in a 3′ splicing site results in the skipping of an exon. **c** A nucleotide substitution leads to the activation of a cryptic splicing site and results in the incorporation of some intronic sequence into the mRNA transcript. **d** A nucleotide substitution deactivates the normal splicing site, while a cryptic one is used a few base-pairs downstream in the intronic sequence. **e** The substitution leads to the skipping of the last exon. All of these situations have been observed after positional cloning of mouse mutations

All types of splicing defects that are theoretically possible have been actually identified in the mouse, altering more or less significantly the function of the encoded protein. A situation that is quite common and has severe consequences is when a 3′ splicing site (3′ss) is altered, leading to the attachment of a stretch of intronic DNA at the 3′ end of the mRNA molecule. In this case a number of amino acid residues are added to the C-terminus of the protein until, by chance, a stop codon occurs to terminate the aberrant transcription. In this case the protein is almost always abnormally folded and accordingly non-functional. Sometimes it also happens that cryptic 3′ or 5′ splice sites are activated after a single point mutation. In this case the consequences are unpredictable although, in general, severe.

Unlike for the splicing sites, mutations affecting DNA binding sites or regulatory regions are not common. This is either because these sites do not represent an important target in which mutations can occur or, alternatively, because mutations occurring at these sites have consequences that are not critical and accordingly are more or less tolerated or compensated for.

Most of the spontaneous mutations which have been found in the mouse, and which have been characterized at the molecular level after positional cloning, have

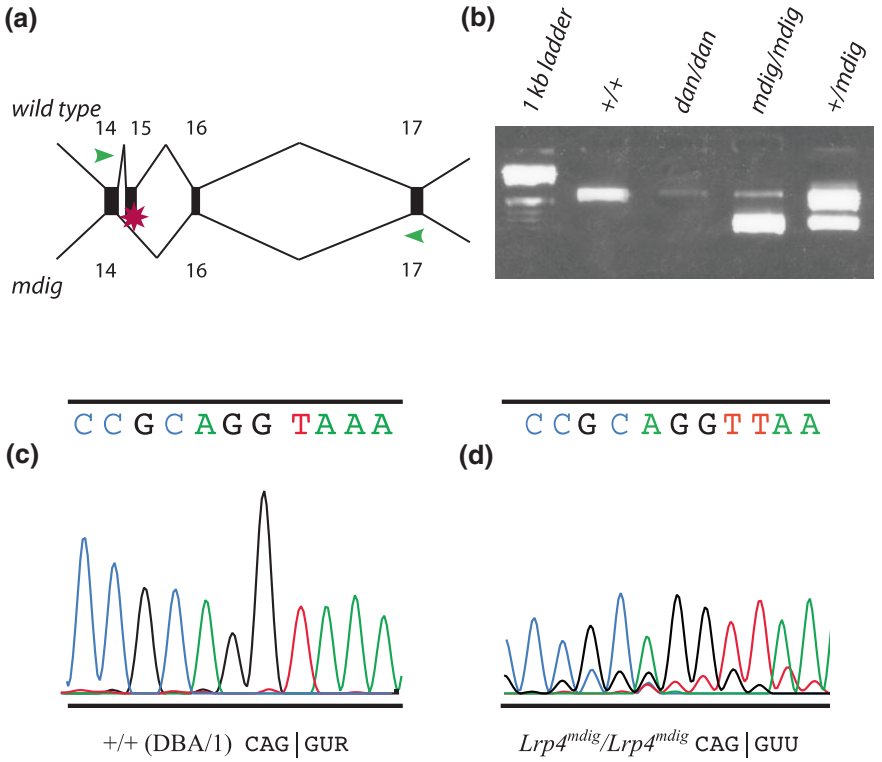


Fig. 7.3 *Mutations resulting in abnormal splicing.* $Lrp4^{mdig}$ and $Lrp4^{dan}$ are two independent recessive mutations affecting the gene encoding the mouse lipoprotein receptor 4 (*Lrp4*-Chr 2). **a** Schematic representation of exons 14–17 of the *Lrp4* gene indicating skipping of exon 15 in $Lrp4^{mdig}/Lrp4^{mdig}$ mice. **b** RT-PCR amplifications performed on total cDNAs with specific primers (green arrows) allow the detection of an amplification product of the expected size in wild type (+/+) whereas only a faint band is observed with $Lrp4^{dan}/Lrp4^{dan}$ cDNA. This is because a retroviral insertion in intron 2 of the $Lrp4^{dan}$ allele hampers the transcription of a messenger RNA. However, the retroviral insertion does not suppress the transcription entirely since a faint band can be observed with cDNAs from homozygous $Lrp4^{dan}/Lrp4^{dan}$. PCR amplification with the same primers yields a product shorter than expected in homozygous $Lrp4^{mdig}/Lrp4^{mdig}$ mice. Here again, skipping of exon 15 is probably not absolute since a faint band is still observable. **c** and **d** Genomic sequence in $Lrp4^{+}/Lrp4^{+}$ and $Lrp4^{mdig}/Lrp4^{mdig}$ co-isogenic mice. An A → T transversion alters the splicing donor site 3′ of exon 15 (from Simon-Chazottes et al. 2006)

been explained by the observation of a non-ambiguous structural defect. Among the few exceptions, one may cite the case of the $Agtpbp1^{pcd}$ allele at the ATP/GTP binding protein 1 locus (formerly known as Purkinje cell degeneration—*pcd*; Chr 13). At this locus six spontaneous alleles and five chemically induced alleles have been reported, which all belong to the same *complementation group* (i.e., they fail to complement each other in a complementation test). For all the mutant alleles, obvious changes have been described in the coding region or splicing sites except for the original $Agtpbp1^{pcd}$ allele. For this allele, Northern blot analysis failed

to detect a transcript in all tissues of homozygotes except for the testis, where reduced levels were noted. In this case, the researchers suggested that the structural defect for this mutation should likely be in a regulatory region. However, as of today, the question is still open (Fernandez-Gonzalez et al. 2002).

With the rapid development of DNA sequencing techniques and the concomitant reduction in costs, it is likely that many regions of the mammalian genomes suspected of having particular importance in the regulation of gene expression will be easily compared between different strains or subspecies. In so doing, many point mutations of potential interest are likely to be discovered outside of splicing sites and regulatory regions. The discovery of a point mutation in the seed region of miRNA96, which is responsible for or associated with the semi-dominant deafness phenotype of *Diminuendo* mice ($Mir96^{Dmdo}$), is a good example and might be the first in a long series of such findings (Lewis et al. 2009).

7.2.3 Insertions, Deletions, and Duplications

Insertions are mutations resulting from the intercalation of a DNA sequence of variable size into the genome. The reciprocal alterations, those that are characterized by a missing sequence or portion of DNA, are called *deletions*. Insertions/deletions can be as small in size as a single nucleotide or, on the other hand, they can expand over several kilobases of DNA, affecting a variable number of genes on a chromosome and sometimes making their analysis difficult.

When aligning DNA sequences in the non-coding regions it is not always easy to select the appropriate designation between insertion and deletion. Sometimes it is noted that a single nucleotide makes a difference, but it is impossible to determine whether the mutation represents an insertion in one of the sequences or a deletion in the other. The situation is even more complex when this single nucleotide difference is frequent and co-localized across different strains. For these cases, geneticists have coined the word *indel* (from insertion/deletion), indicating their ignorance concerning the historical sequence of the molecular change and the co-existence of the two forms as alleles. In short, an indel is a gain or loss in nucleotides, at a specific site, that is polymorphic in a given species.

Microindels are indels that result in a net gain or loss of 1–50 nucleotides. Insertions, deletions, and indels are potentially innumerable since nucleotides can be either deleted or inserted almost anywhere in a DNA strand as a consequence of aberrant replication, unequal crossing-over or transposition. Interestingly, however, deletions are more commonly observed in practice than insertions in both the mouse and human genomes (17 % versus 6.4 %, respectively).

Deletions or insertions of nucleotides have consequences when they occur in an open reading frame (ORF), in the close vicinity of splicing sites or at DNA binding sites. When they occur in an ORF and have a size greater or less than three nucleotides (i.e., a number not divisible by three), they result in frameshift mutations, whose effects are similar to those of the mutations occurring in the splicing sites

and are transcribed, in general, into aberrant mRNA molecules (Perez et al. 2013). When indels have a size of three or a multiple of three nucleotides, they result in the incorporation of additional amino acids into the protein chain, and their effects are difficult to predict. One such example has been described for another allele at the same *Agtpbp1* locus (already mentioned above), the *Agtpbp1*$^{pcd\text{-}5J}$ allele of spontaneous origin. Positional cloning of this mutation demonstrated that, in this case, a GAC triplet was inserted at position 775, adding an additional aspartic acid (Asp) to the protein. Northern blotting demonstrated comparable expression to that of wild-type mice, indicating normal RNA expression. However, Western blot analysis showed that the protein level is dramatically reduced (Chakrabarti et al. 2006).

Many mouse mutations of spontaneous origin, or discovered via studies of the effects of radiation on the germline, are the consequence of deletions encompassing several contiguous genes. Although common, this type of mutation is of limited interest for modeling human defects or even for annotating the mouse genome, because it is in general difficult to establish a direct link between a particular phenotypic trait and the genotypic defect. The mouse mutation *oligotriche* (*olt*-Chr 9) is an example of such a deletion. This mutation has been found to be a 234-kb deletion affecting no less than six contiguous genes: *Vill*, *Plcd1*, *Dlec1*, *Acaa1b*, and parts of *Ctdspl* and *Slc22a14,* but the gross phenotypic expression is relatively modest: some hair loss on the hind legs and male sterility due to severe sperm defects (Runkel et al. 2012).

Duplications are another type of mutation whose effects and consequences are similar to insertions. The gene encoding the leptin receptor (*Lepr*-Chr 4), with all its many alleles, is a good example illustrating both indels and duplications (see Fig. 7.4).

7.2.4 Triplet Expansions

Triplet expansion or trinucleotide expansion is a defect in DNA replication that is responsible for a dozen severe human diseases (i.e., Huntington disease, Friedreich ataxia, X-fragile syndrome, Kennedy syndrome, Steinert myotonic dystrophy, to mention just a few). These diseases are characterized at the DNA level by an increase in the number of tandemly repeated specific trinucleotides, for example CGG, CTG, CAA or CAG, occurring in specific genes and caused by slippage during DNA replication. Huntington disease (HD), for example, is caused by the expansion of CAG repeats in the gene encoding huntingtin (*HTT*). The number of CAG repeats increases with age in some patients and encodes an expanded glutamine (Gln) tract within the huntingtin protein. When the number of repeats passes a critical number (actually 36 for *HTT*), then the enlarged polyglutamine fragment in the protein leads to the formation of the huntingtin aggregates that are observed in the brain as well as in some other tissues, leading to severe pathologies.

Although the mechanism leading to triplet expansion is only poorly understood, geneticists have established that the number of repeats is frequently variable from

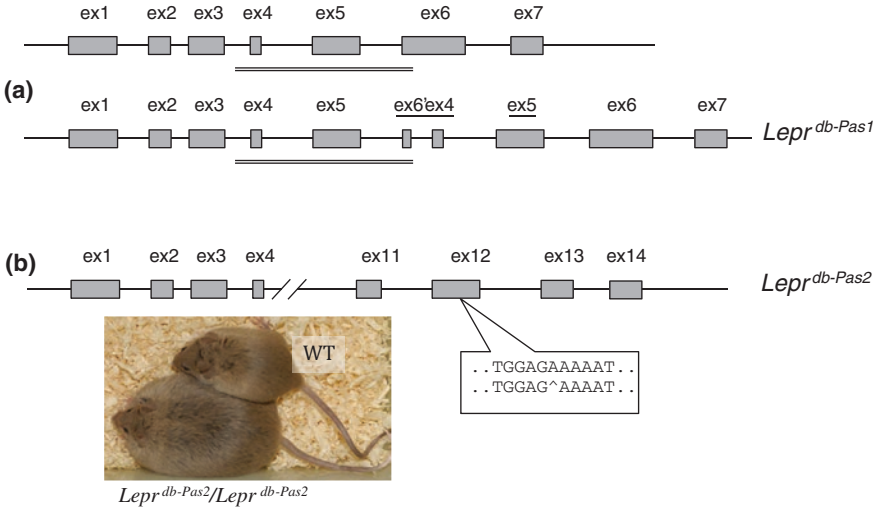


Fig. 7.4 *Mutations resulting from duplications and deletions*. In the mouse, over 15 spontaneous mutations have been reported at the locus of the gene encoding the leptin receptor (*Lepr*-Chr 4). This gene normally consists of 18 exons and has multiple splice variants, comprising at least five isoforms. **a** Among these mutant alleles, $Lepr^{db\text{-}Pas1}$ is the consequence of a partial duplication that spans the entirety of exons 4 and 5, plus 21 bp of coding exon 6 (as well as the two introns between exons 4 and 6). This produces a null allele that is unable to encode a functional receptor (from Liu et al. 1998). **b** Another spontaneous allele, $Lepr^{db\text{-}Pas2}$, is the consequence of a 1-bp deletion producing a frameshift in exon 12, altering another domain of the protein. The mutant allele is inactive and the mouse becomes obese and diabetic

tissue to tissue in the same patient, suggesting that distinct expansion processes can occur in different tissues. Human geneticists have also established strong correlations between the length of the triplet repeats and the severity of the disease.

Such spontaneous cases of trinucleotide expansions have not been reported in the mouse but mouse models of HD, displaying phenotypes relevant to the human disease, have been created by transgenesis (Menalled and Chesselet 2002). These models will aid the understanding of the fundamental mechanisms underlying unstable triplet expansion in humans, and hopefully will also provide useful targets for inhibiting disease development.

7.2.5 *Mutations Resulting from the Insertion of Mobile Elements*

As we discussed in Chap. 6, many mobile elements (retrotransposons, retroviruses, LINEs, SINEs, etc.) are well-known and quantitatively important components of the mouse genome. These elements move within the genome by duplication and retrotransposition and, depending on their integration site, they may have a mutagenic

action. Many such mutations have been identified in the mouse. For example, the *dilute* (*Myo5a*^{d}-Chr 9) mutation, a very ancient mutation of the mouse with several alleles, is the result of the integration of the ecotropic murine leukemia virus *Emv-3* into the myosin VA (*Myo5a*) gene. The *a* (non-agouti-Chr 2) mutation is also the consequence of the insertion of a 5.5-kb virus-like element (VL30) into the first intron of the *agouti* gene, which interferes with the transcription process. At the same *Agouti* locus, we previously reported the case of the dominant mutation A^{vy} (viable yellow), which is the consequence of the insertion of an intra-cisternal A-particle (IAP or retrotransposon) into a non-coding exon at the 5′ end of the *agouti* gene. Similarly, the spontaneous mutation *spastic* (*Glrb*^{spa}-Chr 3) results from the insertion of a 7.1-kb LINE-1 element within intron 6 of the gene encoding the glycine receptor, beta subunit (Mülhardt et al. 1994). Finally, the *hairless* (*Hr*^{hr}-Chr 14) mutation in mice was caused by the insertion of a murine leukemia virus into intron 6 that results in aberrant splicing of the *Hr* gene (Stoye et al. 1988). Some strategies have been designed to make use of the capacity of transposons to move in the mammalian genome, for the induction of new mutations in the mouse and mostly in the rat. We will come back to this point later in this chapter (Sect. 7.6).

7.2.6 Mutations Due to Non-homologous Recombination or Non-homologous End Joining

Non-homologous DNA recombination or non-homologous end joining (NHEJ) occur in mammalian genomes when double-strand DNA breaks are imprecisely repaired, leading to loss (or duplication) of a segment of nucleotides. Spontaneous mutations that are the consequence of NHEJ have been reported in humans (for example, a β-thalassemia leading to hemoglobin Lepore syndrome). To date, no such mutations have been reported in the mouse, although they may theoretically occur spontaneously. However, the NHEJ DNA repair mechanism, along with homologous recombination, is the molecular basis of new genome editing technologies with engineered nucleases (this point will be discussed in Chap. 8).

7.2.7 Copy Number Variations

As already discussed in Chap. 6, structural changes that result in copy number variations (CNVs) in a specific chromosomal region are common in all genomes. In the mouse, approximately 100 genomic regions across the 19 autosomes have been shown to harbor CNVs, ranging in size from 20 kb to 2 Mb, with more than 90 % sequence conservation. These CNVs may be considered to be mutations of a new class: the "*multi-duplications*". They certainly affect gene expression by altering the transcript dosage and, accordingly, the phenotypic variability in genetic diseases by affecting the penetrance of the trait (Cutler and Kassner 2008). CNVs probably play an important role in quantitative genetics.

7.3 Spontaneous Mutation Rates

Spontaneous mutation rates are difficult to assess accurately in any mammalian species for a number of reasons. First, it is clear that only a fraction of the mutations are detectable at the phenotypic level, and this fraction fluctuates from one locus to the next. For example, dominant alleles resulting in lethality in utero or shortly after birth and those impairing the reproductive capacities of animals are often not even identified as heritable traits. Another major difficulty in the detection of mutations is that some of them frequently exhibit wide variations in expressivity or have a very subtle phenotypic expression and accordingly are underevaluated. Recessive mutations are easier to detect because they are in general observed recurrently, especially when they occur in an inbred strain but, even in this case, many mutations have not been identified simply because they have a late onset or because they are expressed only in some particular conditions. For example, most inbred mouse strains are susceptible to experimental infections with flaviviruses (yellow fever or dengue, for example) while a few others are resistant. This susceptibility is caused by a recessive mutation (*Oas1b* locus-Chr 5) and was discovered incidentally during an experiment, but the mice of both strains (resistant and susceptible) look perfectly "normal" for all other characteristics and, for this reason, the mutation remained undetected for many years. Similarly, some strains are susceptible to the antiparasitic drug ivermectin while most others are resistant but, here again, the mutation is cryptic, conditional, and can be detected only when the drug is administered. In conclusion, one can say that the identification of a spontaneous mutant phenotype depends upon the quality and accuracy of the phenotyping and, as a consequence of this, one must bear in mind that mutation rates are, in general, underestimated unless they are computed at a specific locus.

The first estimation concerning the mutation rate towards a recessive allele was published in 1966 by two scientists at The Jackson Laboratory (Schlager and Dickie 1966, 1967). Their estimations were established from the observation of 1,349,725 interstrain F1 progeny at five specific coat-color loci (non-agouti, *a*; brown, formerly *b* and now $Tyrp1^{b}$; albino, formerly *c* and now Tyr^{c}; dilute, formerly *d* and now $Myo5a^{d}$; and leaden, formerly *ln* and now $Mlph^{ln}$).[10] The authors reported 12 mutations from the wild-type allele towards a recessive allele (forward mutations) and calculated an average mutation rate of 8.9×10^{-6} per locus per gamete (with $4.6–15.5 \times 10^{-6}$ as 95 % confidence limits).

Another estimation based on similar crosses, although between different strains, was published a few years later by (Russell and Russell 1996). A total of 1,485,036 F1 progeny were scored at seven loci (the same five as above except

[10] Computations of the mutation rates were made on several interstrain F1 hybrids expected to be all heterozygous for one or several of the recessive coat color alleles and the corresponding wild-type allele. In such an F1 population, the mice with a non-wild-type phenotype are potential carriers of a new mutant allele. This was confirmed by setting up separate crosses.

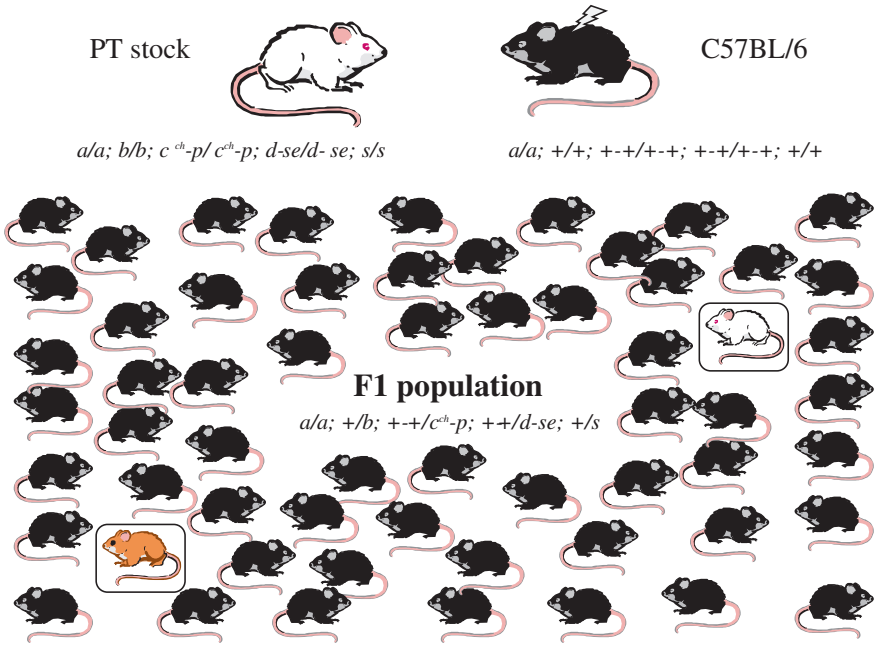


Fig. 7.5 *Assessing the mutation rate at specific loci*. Mice of the PT stock are homozygous for seven recessive mutant alleles involved in the determinism of coat color. When crossed with mice of the C57BL/6 inbred strain (which are non-agouti *a/a* and homozygous for the wild type allele at the other six loci), all F1 are expected to have a non-agouti (*a/a* = *solid black*) coat color phenotype. Phenodeviants, with a coat color different from the expected one (*boxed*), are potentially heterozygous for a new recessive allele at one of the six loci of the PT stock, and their status must be characterized by additional crosses. This historical PT stock, developed at Oak Ridge by W. Russell and colleagues, has been extensively used for assessing the mutagenic activity of radiation or chemical compounds. Another similar stock, the HT stock, with different alleles has been developed at MRC Harwell

leaden *Mlph*^{ln}, plus pink-eyed dilution *Oca2*^{p}, piebald *Ednrb*^{s}, and short ear *Bmp5*^{se}) and the authors calculated a rate of 6.6×10^{-6} mutations per locus per generation.[11] In addition to the "complete" mutations, the same authors also found several "mosaic" mutations at five loci, which led them to calculate a corrected mutation rate of 11×10^{-6} per locus per generation (Fig. 7.5).

These mutation rates, calculated independently, are relatively close to each other and definitely represent a good estimation for the loci described above. However, this rate (~10×10^{-6} mutation per locus per generation) is certainly not representative of the "average" mouse locus because the same scientists at The Jackson Laboratory reported a total of only 28 recessive mutations at 26 different loci from a total of 83,368,463 mice examined, yielding an overall spontaneous

[11] These observations were made on the F1 progeny of a cross between a tester stock, known as PT stock, homozygous for seven fully penetrant recessive alleles, and mice homozygous for the wild-type alleles at the same seven loci.

recessive mutation rate of 6.7×10^{-7} per locus per gamete (95 % confidence limits: 5.1–8.7×10^{-7}). This rate, which is only 1/13th of the rate calculated for the forward mutations at the five/seven specific coat-color loci, is probably a better estimate of the overall spontaneous mutation rate towards a recessive allele in the mouse. This was confirmed by scientists working at Harwell using an independent tester stock, the so-called HT stock, homozygous for six recessive alleles with only one recessive allele (non-agouti *a*) in common with the PT stock.

Schlager and Dickie also recorded the number of mutations towards a dominant allele. They collected this information by observing breeding colonies during a 3-year period (36 mutations were collected from a total of 67,161,745 mice), yielding an estimated spontaneous mutation rate of 0.54×10^{-6} per locus per gamete, with 95 % confidence limits of 0.38–0.74×10^{-6} (Schlager and Dickie 1967).

A careful analysis of the mutations (both recessive and dominant) collected by the scientists at The Jackson Laboratory indicated that there are great differences in the mutation rates at the different loci. As we already mentioned, this is certainly a consequence of the fact that many mutant alleles escape detection either because of their unobtrusive (or very severe!) phenotype or late onset phenotype. This may also be explained by differences in the size of the different loci at the DNA level or the splitting of the coding regions into many exons, offering a wider target to the mutagenic events. However, these two explanations are clearly not sufficient to explain some of the observed differences, and it is now well established that some genes have an unexpectedly higher mutation rate than average. This is the case, for example, with the gene encoding the *Kit* receptor tyrosine kinase (*Kit*-Chr 5), in which 18 spontaneous mutant alleles were recorded in a population of mice analyzed by Schlager and Dickie during their survey.[12] This is also the case with a locus on chromosome 4, where no less than seven independent mutations were found in a single experiment (Kiernan et al. 2002). Other examples are the non-agouti locus (*a*-Chr 2) with 58 spontaneous alleles, and the dilute locus (*Myo5a*-Chr 9 with 53 alleles. Regardless of the loci and observed variations in the mutation rates, these rates remain very low. This explains why mammalian geneticists, like other geneticists, have invested in the development of strategies to increase the rates of mutation.

7.4 Mutagenesis in the Mouse

Over the last century, mice have been extensively used by geneticists as "living test tubes" for assessing the genetic hazards associated with the domestic use of nuclear energy. Mice have also been used by toxicologists for assessing the mutagenic activity of potentially hazardous chemical compounds in the human environment (drugs, food additives, pollutants, pesticides, etc.), and hundreds of mutations of all types have been

[12] On a total of 36 dominant mutations.

produced as "by-products" of these activities. These mutations, in addition to the spontaneous mutations that were previously collected at low frequency in breeding facilities, have been instrumental for the development of mouse genetic maps because, at that time, they were the only available genetic markers. They also provided geneticists with many potentially interesting models of human diseases. However, experimental mutagenesis sensu stricto, which means the treatment of animals with known mutagenic agents to purposefully increase the mutation rate, is only recent.

7.4.1 Gametogenesis and Experimental Mutagenesis

Experimental mutagenesis consists of exposing progenitors of either sex to a mutagenic agent, with the aim of increasing the occurrence of novel mutations in the progenies of the treated animals. For practical reasons, only male progenitors are exposed to the mutagens because spermatogenesis is a continuous process, starting at puberty and lasting several months or even years. In females, on the contrary, gametogenesis is a cyclic process and the number of cells that are potential targets for mutagenesis is much reduced in adult mice (Fig. 7.6).

Spermatogonia are the stem cells of the male germline. When they divide they produce two daughter cells: a spermatogonium type A_0, which stays in the pool of

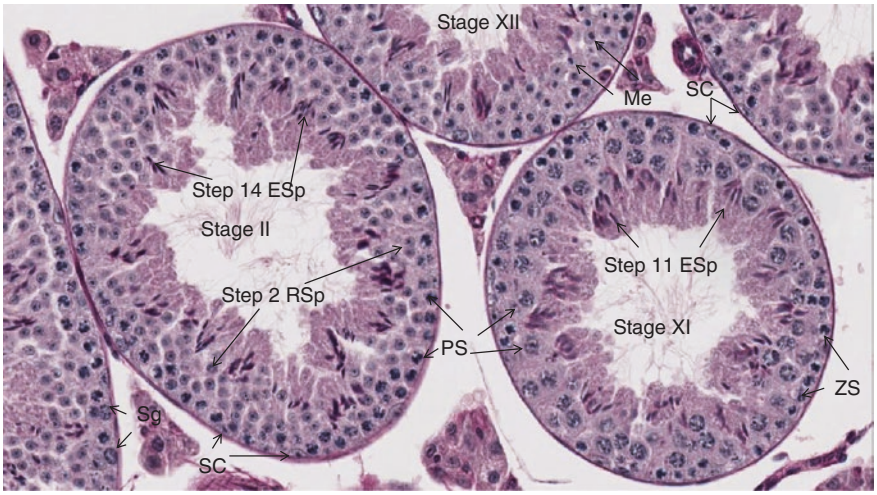


Fig. 7.6 *Histological appearance of seminiferous epithelium*. Sections from an adult mouse testis indicating several stages of the spermatogenetic process. SC = Sertoli cell, Sg = spermatogonia, ZS = zygotene spermatocytes, PS = pachytene spermatocytes, Me = meiotically dividing spermatocytes, step 2 RSp = step 2 round spermatids, step 11 Esp = step 11 elongating spermatids, step 15 Esp = step 14 elongating spermatids, stage II, stage XI, and stage XII = tubules in different stages of the spermatogenic cycle (Figure *courtesy* of Dr. Dianne Creasy, Huntingdon Life Sciences, East Millstone, NJ, USA)

stem cells, and a spermatogonium type A_1 that undergoes several mitotic rounds, producing A_2, A_3, A_4, and Intermediate types, and finally type B spermatogonia. The type B spermatogonia divide and form pre-leptotene spermatocytes, which are almost identical to type B spermatogonia in appearance, but they become much larger as they duplicate their chromosomes to form tetraploid cells and proceed through meiotic prophase (zygotene, pachytene, diplotene, and diakinesis). The first meiotic division produces two short-lived diploid secondary spermatocytes, which rapidly divide again (second meiotic division) to produce four round haploid spermatids. These round spermatids then undergo a complex morphological transformation into spermatozoa, developing condensed heads covered by an acrosome and attached to a motile tail, which are then shed into the tubular lumen (spermiation). In theory, a single A_1 spermatogonium would give rise to 256 sperm cells in 5 weeks, but there is some attrition of cells during spermatogenesis so that the actual number of sperm is smaller than the theoretical maximum. A few of these mature sperm cells will fertilize ova, and most others are eliminated while a new cycle of spermatogenesis follows. The duration of the spermatogenetic cycle is much shorter in the mouse than in most other species; spermatogonia become mature spermatids that are released into the lumen in only 5 weeks (Russell et al. 1990). By comparison, the spermatogenic cycle is 8 weeks in the rat and 10 weeks in humans. It then takes another 1–2 weeks for the released sperm to reach the tail of the epididymis, where they are stored prior to ejaculation.

Mutagenic agents (physical or chemical) exert their effects as soon as they are in contact with the genetic material of the treated mice and this effect terminates, in general, immediately or shortly after treatment ends. The cells that have been mutagenized repair most of the damage resulting from the treatment, but, depending on the severity of this damage, some cells may recover and pass genetic alterations to their daughter cells while others die and are eliminated. The success of a mutagenic treatment is reflected in the percentage of cells that survive and carry a mutation, and the higher this percentage the better. As we will discuss, this depends upon the mutagenic treatment, the type of cells exposed to the mutagen, the dose and duration of the treatment, and the dose rate and the possible splitting of the dose.

Because spermatogenesis is a continuous and precisely timed process, we can calculate the precise stage of development of a specific germ cell, at the time of exposure to a mutagen, depending on the time elapsed between the treatment and the fertile mating. For example, if male mice are exposed to a mutagen and mated 3–4 weeks later, the embryos that result from the mating will have originated from germ cells that were mature spermatids (post-meiotic stage) or spermatozoa entering the epididymis at the time of treatment. In contrast, if the mating takes place more than 7 weeks after the treatment, the embryos result from cells that were exposed as spermatogonia. When the stem cells of spermatogenesis (i.e., the spermatogonia A_0) are successfully mutagenized, the male becomes a permanent provider of mutations. On the other hand, when the targeted cells are post-meiotic (spermatids or spermatozoa), the mutagenesis is transient.

An important point to mention is that a very efficient selection process operates during gametogenesis to eliminate the mutations that may have occurred either

spontaneously or after the mutagenic treatment. This process is much more efficient during the early (diploid) phases of gametogenesis, where the cells divide and have an active metabolism with efficient DNA repair mechanisms, than during the haploid phase, when the cells differentiate but no longer undergo mitosis. In the same way, meiosis occurring at the spermatocyte stage is an efficient filter to eliminate the chromosomal rearrangements that interfere with the normal distribution of chromosomes in the daughter cells. Reciprocal translocations or inversions, for example, are strongly counter-selected when they occur in spermatogonia, whereas many of them are transmitted to the offspring when induced in early spermatids.

When males receive a mutagenic treatment, the number of affected stem cells depends on the dose. If the dose is elevated, most spermatogonia are killed and the male becomes permanently sterile. Conversely, if the dose is too low, the lethal effect is limited but the mutation rate is low and the experiment might not be successful. Selecting the best dose is very important and may require preliminary experiments.

7.4.2 The Induction of Mutations by Radiation

Hermann Muller (1927) was among the first to report that X-rays can cause mutations and chromosomal damage in *Drosophila* flies. However, most of the knowledge geneticists have gathered concerning the mutagenic effects of radiations in the mouse results from research conducted at MRC Harwell in England and at Oak Ridge National Laboratory in the United States. An excellent review of these fundamental studies, which may still be useful, can be accessed online in the book "*Biology of the Laboratory Mouse*" in a chapter by Green and Roderick (1966).

In short, we can say that all types of radiation are mutagenic, provided they have sufficient energy to come into contact with the genetic material. Cosmic radiation, a mixture of photons and high-energy protons originating from outer space, constantly showers on all living organisms and is probably responsible of many "spontaneous" mutations. In contrast, UV radiation, consisting of photons with a wavelength between 100 and 400 nm, is mutagenic (and carcinogenic!) only for the cells of the epidermis. Their energy is insufficient to reach the gonads, and accordingly their impact on the genetic material of mammalian species is virtually nil.

Countless experiments have been performed to understand the mutagenic effects of electromagnetic (X- and γ-rays) and corpuscular (protons and β-particles) radiation. These types of radiation are mutagenic because they have a direct effect on the chromosomes and DNA strands; they produce breakages or deletions that are more or less efficiently repaired, depending on the extent of the damage and the efficiency of the repair mechanisms. They are also mutagenic because they produce ionization as they dissipate their energy into living matter, producing a very large number of hydroxyl and hydroperoxyl free radicals that are highly reactive and diffusible elements. From the experiments conducted by health physicists between 1950 and 1970, it was concluded that the mutagenic activity

and the type of mutations produced by exposure to radiation depend on a physical parameter known as *linear energy transfer* (LET), and, of course, on the dose distributed, the duration of exposure, and whether the dose is fractionated. Heavy particles like protons or α- particles have a very high LET and dissipate their energy over a short distance while passing through living matter. Accordingly, they exhibit high mutagenic activity and produce extensive chromosomal breakage. On the other hand, photons such as X- and γ-rays have a much lower LET and are much less mutagenic, producing mostly point mutations or small-sized deletions. For X- and γ-rays, the rate of induced mutations varies linearly with the dose from 0 to 7 grays (abbreviated Gy).[13] Beyond 7 Gy repair mechanisms are saturated, and many cells are affected by several mutations and die.

When a dose of X- or γ-rays is distributed over a short period of time (at high dose-rate), the mutagenic effect of the radiation is more intense compared to the same dose distributed over a longer period of time. Similarly, a single dose of radiation is more damaging to the genetic material than the same dose split into several sessions. This is a consequence of the fact that the DNA repair mechanisms are saturated when the dose is delivered over a short period of time. Males, whose germ cells are constantly in mitotic activity (from puberty until death), are more susceptible to the mutagenic effects of radiation than females, whose germ cells are resting at the time of birth.

All germ cells are sensitive to radiation, but haploid cells (post-meiotic stages, with n chromosomes) are more sensitive than spermatogonia (2n) because the DNA repair mechanisms are almost inactive in these highly differentiated cells. As we already mentioned, mutations induced in spermatogonia may be transmitted to the offspring throughout the life of the mutagenized animal, whereas mutations affecting the post-meiotic haploid cells are transmitted only during the short lifespan of these cells (3 weeks), provided that they fertilize an oocyte.

In mice, the mutation rate after exposure of spermatogonia to 10 Gy of X-rays at 0.9 Gy/min, split into two doses of 5 Gy distributed 24 h apart, was reported to be ~ 500×10^{-6} per locus per gamete, compared to the spontaneous rate of ~10×10^{-6} as mentioned above (Russell 1962, 1963). This mutation rate seems to be the highest possible for X-and γ-rays (~50× the spontaneous rate). This increase in mutation rate due to the splitting of the dose suggests that the first irradiation imposes some sort of synchronization and enhances mutability of the cells, while the second dose yields more mutations than otherwise expected. One could, in theory, obtain a higher frequency of point mutations by using neutrons. However, most mutations produced by this type of radiation are deletions that are frequently incompatible with the survival of heterozygotes. Deletions are also difficult to analyze in the molecular context, in particular when they encompass more than one gene, as is often the case.

[13] Since 1970, the *gray* (*Gy*) has replaced the *rad* as a unit of absorbed radiation in terms of energy per unit of mass. One gray corresponds to one joule of energy absorbed per kilogram of living matter. One Gy is equal to 100 rads.

7.4.3 The Induction of Mutations by Chemicals

Studies of the mutagenic activities of chemicals were initiated by C. Auerbach (Auerbach and Robson 1946; Auerbach 1962), who first reported that 1,1′-thiobis[2-chloroethane], a chemical warfare agent known as "mustard gas" and used during World War I, could cause mutations in *Drosophila* flies. Since these initial studies, toxicologists have identified a large number of chemicals with mutagenic activity. Identification of such molecules has been rationalized by the introduction of laboratory tests using bacteria (for example, the Ames test developed in the 1970s; Ames et al. 1973). More recently, transgenic mice with several copies of bacterial genes integrated into their genome have been developed as tools for mutation assays; for example, the *lac*I model, commercially available as the Stratagene Big Blue® mouse, and the *lac*Z model, available as the Muta™Mouse (Wahnschaffe et al. 2005a, b). These tests, which are very sensitive, relatively inexpensive, and simple to use, allowed the establishment of a very long (and ever-increasing) list of substances with demonstrated mutagenic activity in mammals.

Chemical mutagens have been classified into four categories based on the type of interaction they have with DNA (Vogel and Rohrborn 1970). The first category includes molecules known as base analogs. Molecules of this category (6-aminopurine, for example) are mistakenly used by bacteria during DNA synthesis, leading to the production of transitions or transversions after replication. However, these substances have not been found to be mutagenic in mammals, probably because the metabolic pathways leading to the synthesis of nucleotides are not exactly the same in mammals and in bacteria.

Intercalating agents represent another important group of mutagens. Examples include acridine orange, ethidium bromide, and proflavine. These molecules insert into the DNA helix and bind covalently to the bases of the two strands, leading to deletions occurring during the next round of replication. As with the base analogs, these substances have little effect on pre-meiotic germ cells of mammals and have not been used frequently for the purpose of experimental mutagenesis. Some of these agents, however, are active on post-meiotic germ cells and induce translocations and deletions at a low rate.

The third class of mutagens includes the deaminating agents that are best represented by nitrous acid and sodium bisulfite. Because deamination of guanine (G) or adenine (A) occurs spontaneously in most eukaryotic cells, it has been suggested that deaminating agents might be mutagenic by increasing the basic, natural level of deamination. This, however, has never been clearly demonstrated in mammals, and these agents are not currently used for mutagenesis.

Alkylating agents are, by far, the most potent mutagens in mammals (mustard gas belongs to this category). Molecules of this type transfer alkyl radicals (methyl, CH_3, or ethyl, C_2H_5) onto DNA bases, particularly on adenine but also on guanine. If this alkylation is not repaired promptly by the DNA repair mechanisms, transitions or transversions ensue during the next step of replication.

Alkylating agents are mutagenic in the mouse, but most of them are only active on post-meiotic germ cells (type-2 spermatocytes or spermatids). Among these substances, we must mention the anticarcinogenic drugs TEPA and Thiotepa™, ethyl methane sulfonate (EMS), methyl methane sulfonate (MMS), triethylenemelamine (TEM), procarbazine, and chlorambucil, all of which have been used during the last thirty years as chemical mutagens in the mouse.

In 1979, William Russell from Oak Ridge National Laboratory reported that a simple alkylating agent, *N*-ethyl-*N*-nitroso-urea (ENU), has considerable mutagenic power, and even more remarkably, that this substance is active on both pre- and post-meiotic germ cells (Russell et al. 1979). These observations had a major impact on genetic research and must be considered as an important milestone in the history of mouse genetics (Fig. 7.7).

ENU is generally sold in the form of a light yellow powder in dark glass bottles, sealed with a rubber stopper. This packaging makes the chemical relatively easy to handle safely. The molecule is light-, heat-, and pH-sensitive and does not dissolve easily in water, but adding a few drops of ethanol avoids this drawback. The mutagenic activity of ENU results from its capacity to transfer an ethyl group to oxygen or nitrogen radicals in the DNA molecule, inducing mis-pairing and ultimately leading to base-pair substitutions or deletions (Van Zeeland et al. 1989; Vogel and Natarajan 1995). In fact, the mutagenic activity of ENU results from two mechanisms acting in opposite directions: the alkylation of the DNA molecule resulting in

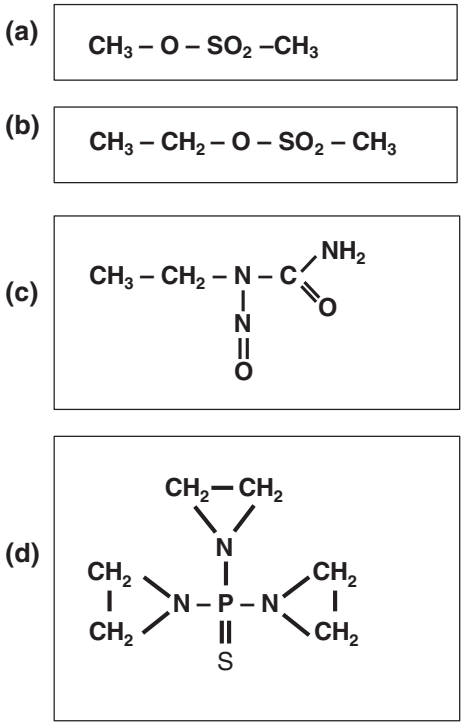


Fig. 7.7 *Alkylating agents with mutagenic activity.* **a** Methyl-methane-sulfonate (MMS). **b** Ethyl-methane-sulfonate (EMS). **c** *N*-Ethyl-*N*-nitrosourea (ENU). **d** *N*, *N*′, *N*″-Triethylenethiophosphoramide (ThioTEPA™). All four molecules are potent mutagens but ENU is, by far, the most potent and the only one active on pre-meiotic germ cells

the creation of adducts on the one hand, and the efficiency of the enzymatic DNA repair mechanisms on the other. In spermatogonia, the ENU-alkylated nitrogen atoms are efficiently repaired, while ENU-alkylated oxygen atoms are repaired with a much lower efficiency.

Many ENU-induced germline mutations have been studied at the molecular level after positional cloning and it has been found that, in the great majority of cases, adenine (A) is the main target of ENU activity with the primary genetic alteration being either AT to TA transversions or AT to GC transitions (Justice et al. 1999).

The mutagenic activity of ENU has been evaluated using several tests (Russell et al. 1979; Favor 1986; Lewis 1991; Lewis et al. 1991, 1992; Favor 1994; Ashby et al. 1997; Schmezer and Eckert 1999). In his initial paper, (Russell et al. 1979) found 35 confirmed mutations at the seven specific loci mentioned above (those homozygous in the PT stock) among 7,584 offspring in the treated group (one injection of 250 mg/kg of body weight), compared to 28 mutations among 531,500 mice in the control group. This indicated a mutation rate 90 times higher than the spontaneous rate and five times higher than for 6 Gy of γ-rays.

Plotting the mutation rates calculated with the same "multiple loci" assay to the doses of ENU injected in male mice, (Favor et al. 1990) observed that the mutation rate for ENU increased roughly linearly with dose, from the threshold dose of ~34 mg/kg of body weight up to 300 mg/kg, a dose that seems to be the highest tolerable by an adult mouse. If the dose remains low, say less than 30 mg/kg of body weight, the mutation rates are not significantly different from the rate of spontaneous mutations in the same assay. Favor's calculations can be summarized in the following two formulae:

$$\begin{aligned} MR \times 10^{-5} &= (1.2 \pm 0.3) \text{ for } D < 33.9 \text{ mg/kg} \\ MR \times 10^{-5} &= (1.2 \pm 0.3) + (0.4 \pm 0.05) \times (D - (33.9 \pm 5.0)) \\ &\quad \text{for } D \geq 33.9 \text{ mg/kg} \end{aligned}$$

where MR = mutation rate and D = dose in mg/kg of body weight.

The threshold effect observed by Favor and colleagues is probably explained by the fact that, when the number of alkylated sites remains low, the repair mechanisms can cope, but when it becomes high or very high, these mechanisms become saturated and mis-pairing increases in proportion to the dose of mutagen.

W. Russell and colleagues reported a few years after their initial publication that three or four injections of 100 mg/kg of body weight, each delivered at weekly intervals, enhanced the mutation rates by a factor 1.8 and 2.2, respectively, compared with a single dose of 250 mg/kg of body weight, while allowing greater survival and fertility of the treated mice (Russell et al. 1982a, b; Hitotsumachi et al. 1985). With such a treatment, the maximum mutation rate of 125–152 $\times 10^{-5}$ per locus could be obtained that roughly corresponds to 150 times the spontaneous mutation rate. It is probably difficult, if not impossible, to increase this mutation rate further because the risk of inducing dominant lethal damage would then be maximized (Fig. 7.8).

This linear dose relationship for induced mutation rates at these seven specific loci demonstrates the extraordinary power of ENU as a mutagen, but cannot adequately predict the absolute rate of induced mutation at an "average" locus in

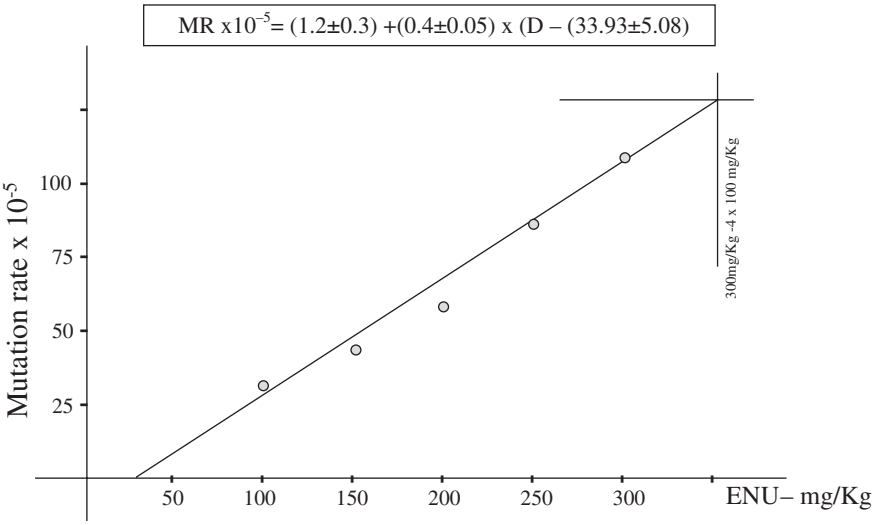


Fig. 7.8 *A dose–response analysis of ethylnitrosourea (ENU)-induced recessive locus-specific mutations in treated spermatogonia*. Predicted locus-specific mutation rates (MR × 10^{-5}) following ENU treatment of spermatogonia in the mouse. This diagram represents the dose–effect linear relationships for the mutagen, between ~34 mg/kg of body weight (the threshold dose under which there is no detectable effect) and 300 mg/kg of body weight, which seems to be the highest dose tolerated by the mouse. This linear model was computed based on extensive data from Neuherberg (Germany) and Oak Ridge (USA) (adapted from Favor et al. 1990)

the mouse genome. Lewis and co-workers, for example, calculated the number of electrophoretic variants induced at 32 loci after treatment with increasing doses of ENU (from 0 to 250 mg/kg of body weight) in DBA/2 and C57BL/6 male mice (Lewis 1991). In these experiments, the mutation rates again appeared to increase linearly with dose but were on average 2.6 times lower than for the "multiple loci" test performed by Russell and colleagues. This latter observation, which has been reported by many other scientists with different tests, indicates that the sensitivity of a locus to the mutagenic activity of ENU probably depends on a variety of parameters such as its "molecular" size, the gene structure (density in A-T, number of introns, etc.), and presumably several other unknown parameters. It is likely that some regions of DNA are more susceptible than others to the mutagenic activity of ENU, validating the idea that hot spots of mutagenesis exist in the mouse genome (Kiernan et al. 2002; Arnold et al. 2012).

The mutation frequency, established by Russell and co-workers for seven specific loci, was later refined by Bode (1984) in another experimental context. Bode considered that, from an optimally mutagenized male, one can expect to obtain, on average, one mutation at a given locus per 1,500 of its gametes. It must, however, be kept in mind that a given male can produce only a limited number of mutations, and this number is dependent on the number of targets that have been hit by the mutagen. From his experimental data, Bode concluded that this number is close to 500 with a dose of 250 mg/kg of mouse body weight. This important

observation means that, when the ultimate goal of an experiment is to produce a great variety of mutations, as is often the case, it is strongly recommended to inject a quite large batch of males rather than to breed many offspring from only a few males.

The mutagenic activity of ENU has been assessed directly at the DNA level in several laboratories, by performing a careful characterization of the number and type of nucleotide substitutions induced, then by matching the nature of these substitutions to the phenotype of the affected mice—if any (Justice et al. 1999; Noveroske et al. 2000; Concepcion et al. 2004; Quwailid et al. 2004; Keays et al. 2006; Takahasi et al. 2007; Arnold et al. 2012).[14] The results of these analyses indicate that ENU induces mutations at a frequency of one for every 0.7–1.9 Mbp of genomic DNA, depending upon the strain and dose. Analysis of the mutations confirms that AT-to-TA transversions occur in about 44 % of the cases while AT-to-GC transitions occur in about 38 % of the cases. When they fall within the coding regions these substitutions cause missense mutations (64 %), splicing defects (26 %) or nonsense mutations (10 %). Another interesting observation, which is a direct consequence of the observed AT-to-TA and AT-to-GC bias mentioned above, is that some amino acid changes are more likely to occur after ENU treatment than others. As such, it must be kept in mind that ENU mutagenesis does not merely increases the spontaneous mutation rate but its action is biased towards certain amino acid changes.

ENU has also been used as a mutagen in the rat and has proved efficient. In this species, however, the dose must be reduced to 90 mg/kg of body weight (Mashimo et al. 2010) and, here again, splitting of the doses has proved more efficient than a single dose.

ENU is a powerful, easy-to-use, inexpensive, and remarkably efficient mutagen. Its effectiveness varies with the strain of mouse treated, and this is why it is absolutely essential to calibrate the experimental parameters as precisely as possible before embarking on a new mutagenesis project. Non-optimal use of the mutagen (i.e., using a dose that is either too high or too low) will inevitably lead to a waste of both time and animal lives.

7.5 Protocols of Experimental Mutagenesis

When a male mouse is treated with a mutagen, for example by performing a single injection of 250 mg ENU per kilogram of body weight, it stays fertile for a few days after the treatment and then becomes sterile for a period spanning 10–18 weeks (Oakberg and Crosthwait 1983). This sterility period is a consequence of spermatogonial cell killing and it is, in large part, strain- and dose-dependent. BTBR, BALB/c, C3H/He, C57BL/6, and DBA/2 strains have been used

[14] The publication by Arnold et al. (2012) is a rich source of information calculated on a very large sample.

for many years, in particular for the large ENU mutagenesis programs conducted in Germany, England, and the USA (Hrabe de Angelis et al. 2000; Nolan et al. 2000; Arnold et al. 2012). These strains appeared to be relatively resistant to ENU, although a relatively higher percentage of C57BL/6 males did not recover fertility after the ENU treatment (Lewis et al. 1991, 1992). Strain FVB, which has several advantages over the other strains for the production of embryos for transgenesis, appeared quite susceptible to ENU, and, accordingly, is not a good choice for experimental mutagenesis (Justice et al. 2000).

While information concerning the toxicity of ENU for the different strains of mice is available, information about the differences in mutation rates is scarce. In an experiment aimed at the production of electrophoretic mutant proteins, Lewis and colleagues (Lewis et al. 1991) made use of C57BL/6 and DBA/2 males, mated to DBA/2 and C57BL/6 females respectively, and did not observe any statistically significant differences in mutation rate between the two strains. Considering the many experiments that have been performed with the classical laboratory inbred strains and the mutagen ENU, one would conclude that, if inter-strain differences in mutation rate were important, this would have been noticed, but this is not the case.

After the sterility period, the spermatogonia that survive ENU treatment progressively repopulate the testis, the sperm concentration rises progressively and the males regain fertility and produce spermatozoa derived from the several different clones of mutagenized spermatogonia. In the sperm population (and later in the embryos), all types of mutations are present but, while dominant mutations can be observed directly in the F1 (or G1) progeny, recessive mutations must be homozygous to express a phenotype. This requires two more generations and the establishment of so-called individual or micro-pedigrees.

The production of mutations in laboratory rodents can be achieved either genome-wide (i.e., at any locus), or in more or less precisely targeted regions, depending on the aim of the experiment and the protocol used. These protocols do not depend upon the mutagen and can apply to radiation as well as to chemicals. We will review the most commonly used mutagenesis strategies.

7.5.1 Phenotype-Driven, Genome-Wide Mutagenesis

A phenotype-driven, genome-wide mutagenesis program consists of four successive steps. First, males (G0) are treated with the mutagen and then mated with females of the same strain, or of another inbred strain, once they have recovered from the sterility period.[15] Unusual phenotypes (often called phenodeviants) that could result from dominant mutations are then looked for by careful examination of the offspring of this first cross (G1 population).

[15] The choice of the strain must be considered with care depending on the future use of the mutant potentially discovered. If mutations are induced, it will definitely be important to identify the background strain in which the mutation occurred.

In the second step, G1 males, which are all potential carriers of recessive mutations at a number of unknown loci, are gathered for the establishment of individual micro-pedigrees. For this, each G1 male is mated to a few females, either of the same or from a different strain, and a sample of six G2 females, offspring of this cross, is selected and crossed (backcrossed in this case) to their father to produce a G3 population. This G3 generation is then carefully examined for the detection of possible recessive mutations. The rigorous and systematic examination of the G3 progeny is part of the phenotyping process and requires much care. Indeed, the higher the number of parameters screened, the higher the number of mutations detected (Fig. 7.9).

Because their deleterious effects are compensated for by the presence of a normal allele in heterozygotes, the recessive mutations induced in the G0 males recur in G3 of the same micro-pedigree and, accordingly, they are easier to detect and preserve than the dominant mutations, which, in most instances, appears only once in the G1 population.

In these micro-pedigrees, when six heterozygous (*+/mut?*) G2 females are backcrossed to the individual G1 males and a minimum of ten G3 offspring are phenotyped per G2 female, the probability of not detecting, just by chance, a

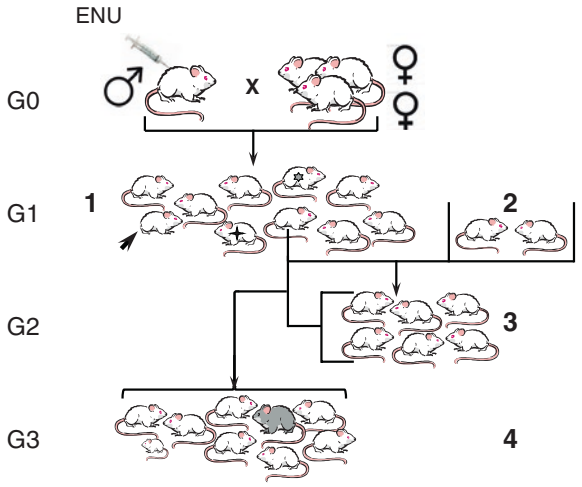


Fig. 7.9 *Phenotype-driven genome-wide mutagenesis*. Phenotype-driven mutagenesis consists of four successive steps. In the first step, males are treated with the powerful mutagen ENU (see text for doses) and mated with 2–3 females after recovery from a 10 to 13-week sterility period (this the G0 generation). The entire G1 progeny is then carefully scrutinized, looking for possible dominant mutations (*arrow*). In the second step, males of the G1 generation (which are potential heterozygous carriers of recessive mutations of all kinds) are selected for the establishment of micro-pedigrees. First, they are mated with females of either the same or a different strain, and 4–6 female offspring (G2) are backcrossed to their G1 father. Finally, the progenies of the G1 male × G2 female offspring (G3) are subjected to careful phenotypic examination (for example, in a "*Mouse Clinic*"). Micro-pedigrees producing mutant phenotypes are then isolated for in-depth analysis. The number of G2 females and their G3 offspring are established after statistical computation to optimize the possibility of detection of new phenotypes

recessive mutation with a visible phenotype that would have been heterozygous in the *+/mut?* G1 males is less than 2 % at the 95 % confidence level.

Bode et al. (1988), followed by McDonald et al. (1994), were among the first to use a whole-genome, phenotype-driven ENU mutagenesis program to produce relevant animal models of phenylketonuria (PKU-OMIM 261640. G0 males were treated with ENU, the G1 male offspring were mated to females of the same strain to produce the G2 progeny, and finally the G1 males and their G2 female offspring were intercrossed to produce the G3 progeny. Blood samples from G1, G2, and G3 mice were analyzed by using the popular Guthrie test, a biochemical test that was used some years ago for detecting elevated levels of phenylalanine in the blood of human newborns.[16] In these experiments, three independent mutant alleles were identified in the G3 populations (*hph1*, *hph2*, and Pah^{hph5}). In addition, it is interesting to note that, using such a phenotype-driven genome-wide strategy, the biochemical pathways at work in the catabolism of the amino acid phenylalanine were literally "dissected" out, with one mutation identified at each biochemical step. This was done in exactly the same manner in which the bacterial geneticists of the early days disentangled the metabolic pathways in bacteria (McDonald 1995).

Nowadays, after much progress in genotyping and phenotyping, several projects have been undertaken by which the G1 and G3 progenies of ENU-mutagenized males have been systematically and extensively phenotyped using a number of criteria by a team of specialists in so-called "*mouse clinics*". Many interesting mutations have been discovered in these projects that would probably not have been noticed in other laboratories (Hoebe and Beutler 2005; Massironi et al. 2006; Arnold et al. 2012). Among the many interesting mutations identified are *Clock*, which modifies the circadian rhythm of affected mice (Wilsbacher et al. 2000), and $Ticam1^{Lps2}$, which results in impaired defense mechanisms against viral and bacterial diseases (Beutler et al. 2007). In a European project comprising six different laboratories and focusing on deafness syndromes, no less than thirteen new independent genes involved in inner ear differentiation and pathology were identified by ENU mutagenesis (Quint and Steel 2003).

The genome-wide production of recessive mutations is a tedious enterprise that requires both intensive animal care and large breeding programs. The advantage of this approach is that no a priori assumptions are made about the genes involved in any pathway. Phenotype-driven mutagenesis is thus an effective method for the identification of novel genes. Numerous projects are now in progress in several laboratories worldwide, where groups of novel mutations, once identified, are roughly phenotyped, mapped to a chromosome, and finally made available to the scientific community for further study. There is no doubt that genome annotation will benefit from all these programs, even if a significant amount of work remains to be achieved after a gene is identified in the form of a mutant allele.

[16] The Guthrie test (a bacterial assay) was routinely used for the neonatal diagnostic of phenylketonuria. It is now replaced either by an immunoassay or by a tandem mass spectrometry assay that measures the amino acid proportions.

7.5.2 The Induction of New Mutant Alleles at Specific Loci

As we remarked above, the main advantage of the genome-wide, phenotype-driven mutagenesis approach is that most of the mutations collected are new alleles appearing at loci where no mutations had ever been isolated before. However, in some cases, it may be desired to induce new alleles at a given locus, for example to explore the possibility that the severity of a given phenotype might be allele-dependent.

The induction of new alleles at a specific locus is well illustrated by the so-called "*multiple loci test*", which was used for the assessment of spontaneous mutation rates, and which we introduced earlier in this chapter. Using this test several new alleles (in particular at the *Tyr*—albino locus) have been induced after treatment with mutagens in the gametes of wild-type male partners, and were then observed directly in the F1 progeny of these males after mating with females homozygous for a set of recessive viable alleles (Rinchik and Carpenter 1999). A similar strategy can be applied to any situation where the production of a series of new alleles at a given locus might be informative, and is ideal when at least one viable recessive allele is available.

Bode (1984) used ENU to produce new alleles at the Brachyury (*T*), quaking (*qk*) and tufted (*tf*) loci by mutagenizing +++/+++ (wild-type) male mice and crossing them to females with the genetic constitution *T qk tf*/++ *tf*. In the F1 progeny of these mice the researcher found three tufted phenotypes [tf], one quaking [qk], and one with a short tail (*t*-interacting or t^{int}) out of 5,172 offspring. In similar experiments, Justice and Bode (1986) and Bode et al. (1988) produced several new alleles at the same three loci, with some of the new alleles at the quaking locus exhibiting interesting and unexpected properties (Justice and Bode 1990; Cox et al. 1999).

Chapman and colleagues performed a similar experiment (Chapman et al. 1989) with the aim of understanding why mice affected by the X-linked Dmd^{mdx} mutation, homologous to the human mutation producing Duchenne muscular dystrophy, were not clinically affected. Chapman hypothesized that this striking phenotypic difference might be because the original Dmd^{mdx} mutation hits a domain of the gene encoding dystrophin that is not functionally essential, and supposed that alleles affecting one of the other four domains of the protein might have more severe effects. To test this hypothesis he created four new alleles at the *Dmd* locus (Dmd^{cv2}, Dmd^{cv3}, Dmd^{cv4}, and Dmd^{cv5}) by ENU mutagenesis. These new mutant alleles were detected by checking for an increase in the creatine phosphokinase (CPK) plasmatic levels in the female progeny of ENU treated +/Y males crossed with Dmd^{mdx}/Dmd^{mdx} homozygous females.[17] The result of these experiments was that all five alleles, the four ENU-induced and the original one, had a very similar

[17] An increase in the plasma level of creatine phosphokinase (CPK) in these F1 mice reveals some damage to the muscular tissue, and is often an indication of the likely occurrence of a new *mdx* allele.

phenotype with no obvious muscular pathology, although the four mutations were found to affect totally different domains. Later, it was demonstrated that mice homozygous for the original *Dmd*^{mdx} allele and one of the ENU-induced series (*Dmd*^{cv5}) had a weaker effect than the other three alleles on the electro-retinogram (ERG) phenotype of the mutant mice (Figs. 7.10 and 7.11).

This observation indicated that the position of the mutation in the dystrophin-encoding gene, although it had no effect on the muscular phenotype, nonetheless had some direct consequences on the ERG phenotype (Pillers et al. 1999). This contributed to the fine annotation of the different domains of the *Dmd* gene, but

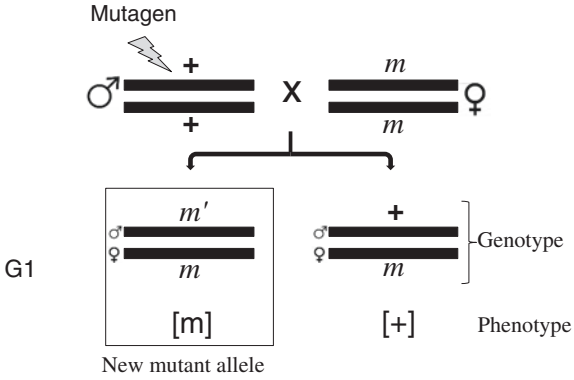


Fig. 7.10 *Targeted chemical mutagenesis*. Male mice are mutagenized and then mated to females homozygous for a recessive allele (*m)* at a specific locus. The G1 offspring of this type of cross are expected to be all wild type. Any deviation from this phenotype must be considered a possible new mutant allele at the *m* locus, especially if some similarities exist between the new phenotype and the phenotype of the female (m). For example, this strategy allowed the generation of an allelic series at the *dystrophin* gene (*Dmd*) (Chapman et al. 1989)

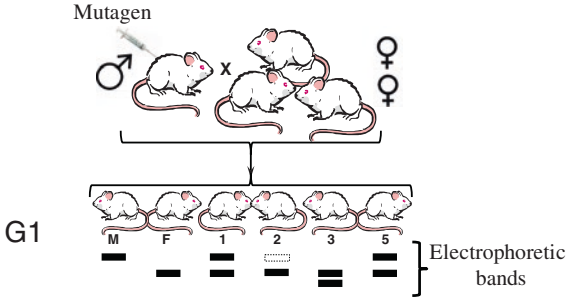


Fig. 7.11 *Targeted chemical mutagenesis*. A male homozygous for a polymorphic protein is treated and then crossed with a female homozygous for another electrophoretic variant, and the G1 progeny are analyzed with the same technique. Mice nos. 1 and 5, as expected, are heterozygous for the two parental forms. Mouse no. 2 is heterozygous for an inactive allele (dotted line) inherited from its (mutagenized) father. Mouse no. 3 is heterozygous for the maternal form and a new functional electrophoretic variant derived from its father

did not explain the phenotypic differences between the human pathology and the mouse model.

A variation of the above-mentioned strategy is to analyze the electrophoretic pattern of enzymatic proteins in an interstrain F1 hybrid where one parent (usually the male) has been mutagenized. Such an "electrophoretic multiple loci test" has been successfully used to identify new mutations at loci encoding for enzymatic proteins (Johnson and Lewis 1981; Marshall et al. 1983; Lewis et al. 1991, 1992).

ENU mutagenesis has also been used to induce mutant alleles in the genes encoding the β-chain of hemoglobin (Peters et al. 1986) as well as to produce several null or functionally different alleles (Charles and Pretsch 1987; Pretsch et al. 1994).

The production of new alleles is also interesting in that it allows the production of slightly different animal models. An excellent example of this situation is provided by the existing animal models of human citrullinemia type I (Perez et al. 2010), where it was demonstrated that some alleles, because they hit a different domain of the protein, appeared to be much better animal models of the human syndrome of citrullinemia (OMIM 215700).

The condition set above—that at least one recessive and viable mutant allele for the locus of interest is available to allow the production of other mutant alleles—is not an absolute prerequisite, and alternative strategies are possible. Let us suppose, for example, that other alleles are desired at the *Mut* locus, which to date has only been characterized by the unviable (or sterile) mutation mut^1. In this case, several F1 (or G1) males, potentially heterozygous for many new ENU-induced mutations (among which is a potentially new mut^2 allele?) are produced and then crossed to $+/mut^1$ females. If, by chance, a mouse with an abnormal [mut] phenotype is detected in the progeny of one of these females, this suggests that a new mut^2 allele at the *Mut* locus has very likely been induced by the treatment. The new allele can then be recovered from the G1 progeny.

7.5.3 The Induction of Mutations in Specific Regions of the Genome

Many strategies have been used to induce and identify the mutations in a specific chromosomal region. Here, we describe three of these strategies that may be of interest in the future: the first makes use of deletions, the second uses consomic or congenic strains, and the last strategy requires a set of overlapping inversions.

Using deletions to detect recessive mutations can only be applied to regions where haploidy is compatible with life. The basic principle is that, when a mutation is induced in the chromosomal segment in front of a deletion, a new phenotype (often lethal) is observed when the chromosome carrying the induced mutation and the deleted chromosome are associated in the same genome. In these conditions, the breeding protocol requires more than one generation, since the induced

mutation must be kept in the heterozygous state while it is revealed by the deletion. The deletion strategy has been used many times (Justice et al. 1997; Rinchik and Carpenter 1999) and has been included in modern mutagenesis programs (Nolan et al. 2000) to identify potential models of human diseases (Fig. 7.12).

The use of consomic strains is an interesting strategy to safely collect the mutations induced in a particular chromosome. Consomic strains (see Chap. 9) are strains in which an entire chromosome has been backcrossed from a donor strain into a different recipient or background strain. Such strains are completely identical for all chromosome pairs but one. These are not common, but at least one set exists (Nadeau et al. 2000), and this is sufficient for the strategy to be applicable. The strategy, presented in Fig. 7.13, is an interesting approach to studying the mutations that have a weak effect or that require sophisticated tests for their detection, because it is possible to establish a co-isogenic strain where the newly induced mutations are safely stored before being studied. This is a great advantage when populations (not only individuals) are to be compared at the phenotypic level; for example, histocompatibility, susceptibility to infectious diseases, and QTL analysis. The same co-isogenic strain that is homozygous for the targeted chromosome can be used several times in successive rounds of mutagenesis experiments, resulting in the progressive accumulation of several new alleles in the targeted chromosome (Fig. 7.13).

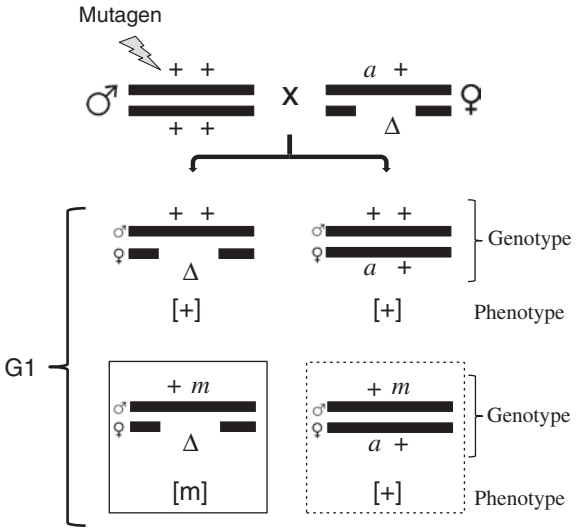


Fig. 7.12 *Using deletions to detect recessive mutations in a specific region.* A male mouse is mutagenized and then mated to females heterozygous for a recessive marker *a*, and for a viable deletion (Δ) (its phenotype is [a]). Most offspring of this cross have a wild-type phenotype, except when a recessive viable mutation *m* is induced by the mutagen in the chromosome segment encompassed by the deletion. In this case, a new phenotype is observed (m). In the cases where the induced mutation is lethal, the progenies are reduced in size and the mice heterozygous for *a* are also heterozygous for *m*, the new mutation (except for a few recombinants)

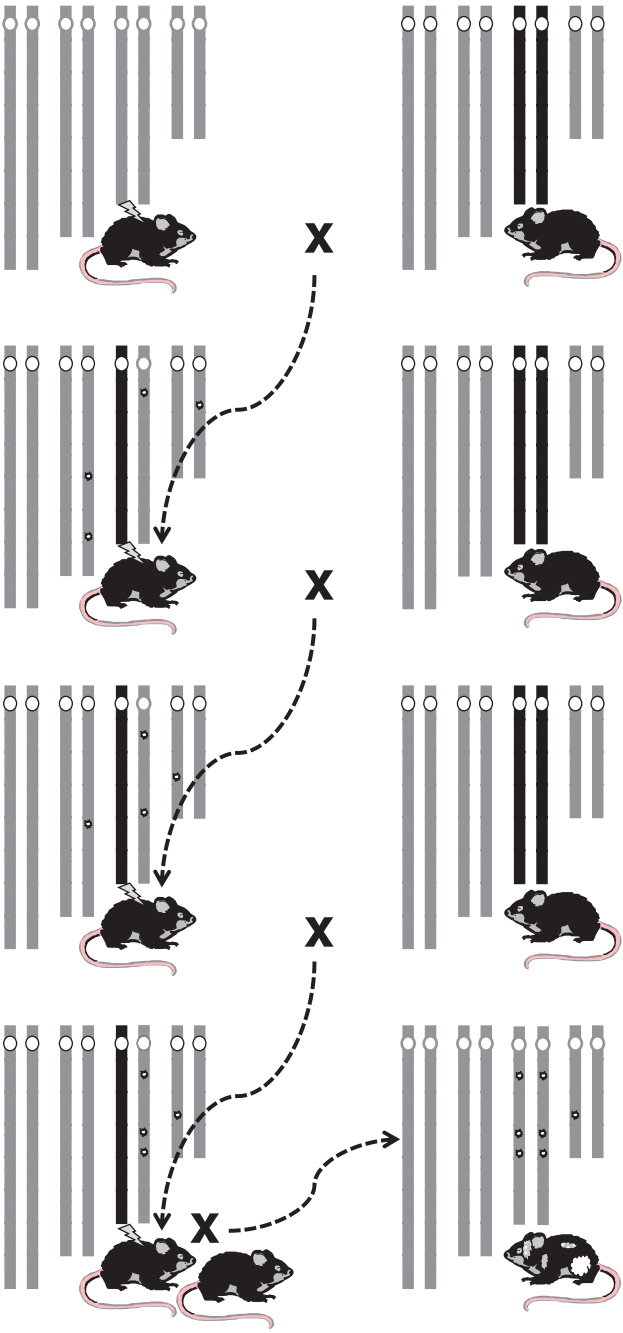


Fig. 7.13 *Accumulating mutations in a specific chromosome*. Male mice are mutagenized and then mated to female mice consomic for a specific (targeted) chromosome (solid black). F1 male offspring of this cross are mutagenized again and crossed to female mice of the same strain, consomic for the same chromosome. The same cycle of mutagenesis—cross with a consomic partner is perpetuated a number of times, and in so doing mutations accumulate only on the targeted chromosome at each generation. Finally a few female offspring of this series of backcross are selected by microsatellite genotyping and backcrossed to their consomic father. This allows re-establishment of a fully co-isogenic strain with many independent mutations accumulated in the same chromosome pair. These mice are then carefully phenotyped. If one of the induced mutations is lethal, the experiment cannot be completed, but the induced mutations can be kept and studied in another context, for example after outcrossing. This strategy requires very little work and a very limited number of animals to be used. This can easily be coupled with a gene-driven mutagenesis experiment, reducing the time spent on genotyping

The use of a set of overlapping inversions is similar in principle to the use of consomic strains and is reminiscent of the balancer chromosome developed in the past by Muller and colleagues for the collection of X-ray induced mutations in *Drosophila melanogaster*. An example of this strategy was the use of a genetically engineered inversion in chromosome 11, which has been described in detail (Zheng et al. 1999; Kile et al. 2003).

Many other strategies have been used to generate and keep mutations in specific areas of the mouse genome that cannot be described in detail here. We will just briefly mention that Shedlovsky et al. (1986, 1988), using a specially designed strategy, were able to induce and study a dozen new lethal alleles within a region spanning two centiMorgans on each side of the *T/t* region on mouse chromosome 17.

7.5.4 A Gene-Driven Strategy for the Production of Mutations at Specific Loci

With the expansion of advanced techniques for the structural analysis of DNA, approaches have been developed that are based on the direct, in vitro detection of DNA alterations, either at specific loci or in specific regions of the genome. These techniques, when applied to the offspring of mutagenized males, allow the production of new mutations in specific regions, ultimately into a preselected (or targeted) gene.

The strategy generally consists of four steps. First, adult males of an inbred strain are treated with an appropriate dose of mutagen (ENU in most instances) and then mated with females of the same inbred strain for the production of a large G1 population.[18] In the second step, sperm samples are collected from adult G1

[18] In this type of experiment it is necessary to exclusively use mice of an inbred strain to enable the non-ambiguous characterization of the mutations potentially induced in the progeny by the mutagen.

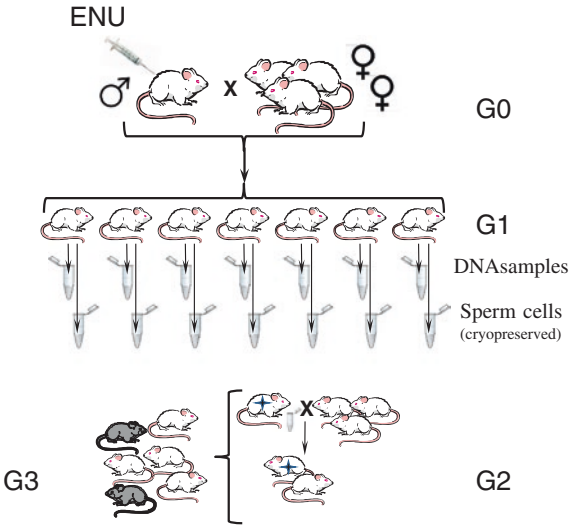


Fig. 7.14 *Genotype-driven mutagenesis*. Male mice are treated with ENU and mated to females (preferably of the same inbred strain) once they have recovered from the sterile period (G0). A large number of G1 males, which all are heterozygous carriers of a great number of independent point mutations (mostly base-pair changes), are then bred. Sperm samples from each G1 mouse are collected and preserved deep-frozen, while DNA samples from the same mice are processed and stored with the same reference. Identification of the mutations generated by the ENU treatment in a specific target (a gene or any other specific sequence) is carried out by molecular techniques to identify DNA mismatches, or directly by sequencing. Once the base-pair changes are identified and considered potentially interesting (stop codons, missense, etc.), the corresponding sperm cells are thawed and heterozygous mice are produced by in vitro fertilization with oocytes of the same background strain. A major advantage of this method is that it produces all types of point mutations, not only knockouts. A drawback is the difficulty of and time required for identifying the mutations in the targeted region. With the rapid expansion of new sequencing techniques, the identification step should be somewhat easier

offspring of this initial cross and stored deep-frozen for performing future in vitro fertilization. Simultaneously, DNA samples from the same G1 males are prepared, cross-referenced with the sperm samples, and stored (Fig. 7.14).

The third step consists of the analysis of the DNA sequence of all G1 mice, looking for any structural changes that may have occurred in a selected and well-delimited region of the genome. This can be achieved by using a sensitive, high-throughput, physical technique, detecting all single nucleotide mismatches after pooling of the DNA samples. This can also be achieved by direct sequencing or SNP genotyping.

When a mutation is found and registered as potentially interesting (i.e., excluding synonymous base-pair changes but retaining nonsense or missense mutations with predicted severe effects), the fourth and last step is performed: the sample of sperm cells corresponding to the potentially interesting mutant mouse is thawed,

oocytes of the same strain are fertilized in vitro and implanted in pseudo-pregnant mothers, and, once born, the potentially heterozygous offspring are bred and crossed in order to produce homozygous offspring whose phenotype is then observed. In this micro-pedigree, the molecular characterization of the offspring is fundamental.

This gene-driven protocol allows the production of all types of mutations (and not only knockouts) in all regions of the genome (coding and non-coding). A drawback is the difficulty and time required for identifying the mutations in the targeted regions. However, with the rapid expansion of modern sequencing techniques, the identification step should be somewhat simplified and shortened in the near future.

The gene-driven or targeted mutagenesis approach has several advantages. It is fast and relatively inexpensive compared to other gene-driven strategies (for example, the engineering of knockouts in ES cells—see Chap. 8). Once identified in a batch of frozen sperm cells, a mutation can be retrieved and made available as heterozygous adult mice in 4–5 months' time. Another interesting point is that a repository comprising a very large number of (non-characterized) mutant alleles can be established by progressively accumulating and storing samples of deep-frozen sperm cells from ENU-treated mice. As we already mentioned, and as observed by direct sequencing of samples prepared from ENU-treated mice, one expects ~0.7–1.9 nucleotide change(s) to be induced per Mbp of mouse DNA after the injection of a single dose of 250 mg/kg. If we consider that the mouse genome consists of 2.7×10^9 bp, one can then expect between ~2,000 and 5,000 de novo substitutions in each G1 progeny from an ENU-treated male mouse. If these nucleotide changes are randomly distributed, one can then expect between ~30 and 75 of the latter to be in the coding DNA or the splicing sites, of which ~25–60 will generate a missense, a nonsense or a splicing defect (77 %).

In addition to these theoretical considerations (but based on actual sequencing data!), one can also calculate that a repository with frozen sperm samples from 20,000 individual G1 animals will be a resource with the potential presence of six independent mutations at any gene of the mouse genome (at the 5 % risk level).

The identification of specific gene alterations can be achieved using pooled DNA samples and run concurrently in several different laboratories to increase the efficiency and ultimately lower the cost of mutagenesis. The final advantage is that, in a species such as mouse where sperm cells can be frozen for long periods and thawed for fertilization, there is no time limit for the identification of mutations. Several laboratories have already published interesting results in this manner (Coghill et al. 2002; Augustin et al. 2005; Michaud et al. 2005; Gondo 2008; Gondo et al. 2010), demonstrating that this gene-driven strategy for the induction of mutations in the mouse might be very promising. This is even truer if we consider that the technique in question can also be applied to the annotation of DNA sequences that are highly conserved across different species; for example, those that are transcribed into non-coding RNAs or not transcribed at all, and whose function is still under scrutiny.

7.6 Other Techniques for the Production of Mutations in the Mouse

In addition to those described earlier in this chapter, a few other strategies have been proposed in the past for the induction of novel mutant alleles in the mouse genome. Most of these techniques have not proved to be significantly more advantageous than the techniques currently in use (ENU mutagenesis in particular) and, for this reason, they have been abandoned. However, exceptions must be made for two strategies that have demonstrated some real advantages. The first consists of treating embryonic stem cells (ES cells) with chemical mutagens (ENU or EMS): this approach will be discussed in the next chapter. The second strategy consists of using transposable elements as insertional mutagens in the mouse, just as the *P* elements were used in *Drosophila melanogaster*, i.e. with the assumption that, when by chance the random insertion of a transposon occurs into a gene, it generally hinders the transcription of a normal mRNA at or near the insertion site and causes a loss-of-function mutation. This technique is known as transposon-based insertional mutagenesis or TIM. We will describe it briefly.

As discussed in Chap. 5, transposable elements (TEs or transposons) are short DNA sequences that move (transpose) within the genome of a great variety of organisms, including bacteria, plants, insects, and vertebrates, by using a cut-and-paste mechanism (i.e., with no RNA intermediate). This mechanism of transposition requires a specific structure of the transposon, with inverted repeats at both ends, and a specific enzyme (a *transposase* or *transposonase*), which is synthesized either by the TE itself (in the case of autonomous transposons) or "*in trans*" by an independent gene (in the case of non-autonomous transposons). Transposons are very active in the genome of plants and bacteria, as well as in some other species, and play an important role in evolution.[19] In mammalian genomes, on the other hand, transposons are inactive and the transposase-encoding genes are degenerated and no longer functional.

Starting from these observations, geneticists had the clever idea to "synthesize" a transposon by genetic engineering using an active transposase in the context of a mammalian genome. To do this, they selected the sequence of a transposon of the *Tc1/mariner* family active in fish (salmon) and, taking into account some phylogenetic data, they could "resurrect" a functional transposon system that they judiciously named *Sleeping Beauty* (*SB10*) in memory of its historical origins. *SB10* was confirmed active in the mouse and rat genomes, inducing mutations by transposition as expected (Ivics et al. 1997).

In experiments making use of the *SB10* transposon system, two transgenic strains are prepared independently, one carrying the transposon proper (sometimes modified to carry a marker cassette that helps track the animal carriers of a novel mutant allele) and the other expressing the indispensable transposase. When

[19] The transposons were discovered and studied in maize by Nobel laureate B. McClintock, precisely because of their mutagenic activity.

desired, the two strains are crossed to generate F1s in which transposition can occur. In the mouse, the frequency of *SB* transposition was estimated to be in the range of 0.2–2.0 events per spermatid (Copeland and Jenkins 2010). Although the rate of production of transposon knockout mutations (TKOs) is less than the rate of mutations resulting from ENU treatment, the TKOs are, in most instances, easier to identify and to clone. By outcrossing the animals carrying the TKO mutations of interest, one can separate the transgene-encoding transposase from the other components of the *SB* system (the mutator element) and transposition immediately stops.

To illustrate the use of transposons as mutagens and the great versatility of this strategy, we recommend a set of interesting publications (Carlson et al. 2003; Lu et al. 2007; Takeda et al. 2007, 2008; Largaespada 2009; Ivics et al. 2011; Furushima et al. 2012). Finally, a review paper by Copeland and Jenkins (2010) is a beautiful illustration of the contribution of the *SB10* system to the analysis of the determinism of cancer and the discovery of cancer genes.

In the mouse, and as we will explain in the next chapter, the transposon *Sleeping Beauty* as well as another one called *piggyBac* have been used extensively both for the transfection and the production of mutations in ES cell lines in vitro.

If transposon-based insertional mutagenesis has some obvious advantages for the production of mutations, it is also interesting for the transfer of genes with stable expression in mouse ES cells. Finally, it may also have applications enabling the persistent expression of therapeutic genes in patients.

7.7 Conclusions

Spontaneous mutations, which are generally identified through the observation of an abnormal phenotype, present several advantages. The first and probably the most important is that they are produced at virtually no cost and are in general freely available. Another advantage is that they have, in general, an obvious phenotype given that they are identified based on observation. Also, spontaneous mutations represent a great variety of molecular events, such as deletions, insertions, and point mutations, generating not only loss-of-function alleles but also hypomorphic and hypermorphic alleles. The problem is that not all mutant genes have an obvious phenotype or, conversely, the phenotype of some mutant alleles is sometimes so severe that affected offspring die in utero.

When mutant allelic forms of a gene are not readily available, the only possible approach for gene annotation is to generate de novo mutations. Thus, the discovery of the extraordinary virtues of ENU as a mutagen can certainly be regarded as a milestone in the history of mouse genetics. With this substance at our disposition, it is now possible to produce and store a great number of new mutant alleles for each protein-coding gene, and all these mutations are a valuable tool for genome annotation. Another advantage is that it is now also possible to induce mutations in those regions of the genome that are highly conserved but whose function is not yet elucidated. The only drawback that should be considered is that mutagens act

randomly, forcing us to make a sometimes lengthy and costly selection among the collected mutations. In this regard, and as we will discuss in the next chapter, the widespread availability of a variety of genetic engineering technologies, including new genome editing tools, has opened the field to the creation of subtle modifications in the mouse genome at will. Even though the identification of genes accountable for single-gene phenotypes is very important, in particular in the context of gene annotation, most of the pathologies that affect human patients are not "monogenic" but are influenced by multiple genes with additive or synergistic effects. As such, our present challenge is to advance the genetic analysis of complex traits.

References

Ames BN, Lee FD, Durston WE (1973) An improved bacterial test system for the detection and classification of mutagens and carcinogens. Proc Ntl Acad Sc USA 70:782–786

Antonarakis SE, Kazazian HH, Gitschier J, Hutter P, de Moerloose P, Morris MA (1995) Molecular etiology of factor VIII deficiency in hemophilia A. Adv Exp Med Biol 386:19–34

Arnold CN, Barnes MJ, Berger M, Blasius AL, Brandl K, Croker B, Crozat K, Du X, Eidenschenk C, Georgel P, Hoebe K, Huang H, Jiang Z, Krebs P, La Vine D, Li X, Lyon S, Moresco EM, Murray AR, Popkin DL, Rutschmann S, Siggs OM, Smart NG, Sun L, Tabeta K, Webster V, Tomisato W, Won S, Xia Y, Xiao N, Beutler B (2012) ENU-induced phenovariance in mice: inferences from 587 mutations. BMC Res 5:577

Ashby J, Gorelick NJ, Shelby MD (1997) Mutation assays in male germ cells from transgenic mice: overview of study and conclusions. Mutat Res 388:111–122

Auerbach C (1962) Mutation: an introduction to research on mutagenesis. Part I: methods. Oliver and Boyd, Edinburgh

Auerbach C, Robson JM (1946) Chemical production of mutations. Nature 157:202

Augustin M, Sedlmeier R, Peters T, Huffstadt U, Kochmann E, Simon D, Schöniger M, Garke-Mayerthaler S, Laufs J, Mayhaus M, Franke S, Klose M, Graupner A, Kurzmann M, Zinser C, Wolf A, Voelkel M, Kellner M, Kilian M, Seelig S, Koppius A, Teubner A, Korthaus D, Nehls M, Wattler S (2005) Efficient and fast targeted production of murine models based on ENU mutagenesis. Mamm Genome 16:405–413

Beier D (2000) Sequence-based analysis of mutagenized mice. Mamm Genome 11:594–597

Beutler B, Du X, Xia Y (2007) Precis on forward genetics in mice. Nat Immunol 8:659–664

Bode VC (1984) Ethylnitrosourea mutagenesis and the isolation of mutant alleles for specific genes located in the T region of mouse chromosome 17. Genetics 108:457–470

Bode VC, McDonald JD, Guénet JL, Simon D (1988) hph-1: a mouse mutant with hereditary hyperphenylalaninemia induced by ethylnitrosourea mutagenesis. Genetics 118:299–305

Carlson CM, Dupuy AJ, Fritz S, Roberg-Perez KJ, Fletcher CF, Largaespada DA (2003) Transposon mutagenesis of the mouse germline. Genetics 165:243–256

Chakrabarti L, Neal JT, Miles M, Martinez RA, Smith AC, Sopher BL, La Spada AR (2006) The Purkinje cell degeneration 5 J mutation is a single amino acid insertion that destabilizes Nna1 protein. Mamm Genome 17:103–110

Chapman VM, Miller DR, Armstrong D, Caskey CT (1989) Recovery of induced mutations for X chromosome-linked muscular dystrophy in mice. Proc Natl Acd Sc USA 86:1292–1296

Charles DJ, Pretsch W (1987) Linear dose-response relationship of erythrocyte enzyme-activity mutations in offspring of ethylnitrosourea-treated mice. Mutat Res 176:81–91

Coghill EL, Hugill A, Parkinson N, Davison C, Glenister P, Clements S, Hunter J, Cox RD, Brown SD (2002) A gene-driven approach to the identification of ENU mutants in the mouse. Nat Genet 30:255–256

Concepcion D, Seburn KL, Wen G, Frankel WN, Hamilton BA (2004) Mutation rate and predicted phenotypic target sizes in ethylnitrosourea-treated mice. Genetics 168:953–959
Copeland NG, Jenkins NA (2010) Harnessing transposons for cancer genes discovery. Nat Rev Cancer 10:696–706
Cox RD, Hugill A, Shedlovsky A, Noveroske JK, Best S, Justice MJ, Lehrach H, Dove WF (1999) Contrasting effects of ENU induced embryonic lethal mutations of the quaking gene. Genomics 57:333–341
Cutler G, Kassner PD (2008) Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. Cytogenet Genome Res 123:297–306
Favor J (1986) The frequency of dominant cataract and recessive specific-locus mutations in mice derived from 80 or 160 mg ethylnitrosourea per kg body weight treated spermatogonia. Mutat Res 162:69–80
Favor J (1994) Specific-locus mutations tests in germ cells of the mouse: an assessment of the screening procedures and the mutational events detected. In: Mattison DR, Olsham AF (eds) Male-mediated developmental toxicity. Plenum Press, New York, pp 23–36
Favor J, Sund M, Neuhauser-Klaus A, Ehling UH (1990) A dose-response analysis of ethylnitrosourea-induced recessive specific-locus mutations in treated spermatogonia of the mouse. Mutat Res 231:47–54
Fernandez-Gonzalez A, La Spada AR, Treadaway J, Higdon JC, Harris BS, Sidman RL, Morgan JI, Zuo J (2002) Purkinje cell degeneration (pcd) phenotypes caused by mutations in the axotomy-induced gene, Nna1. Science 295:1904–1906
Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, Pethiyagoda CL, Stuve LL, Johnson FM, Daly MJ, Wade CM, Cox DR (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448:1050–1053
Furushima K, Jang CW, Chen DW, Xiao N, Overbeek PA, Behringer RR (2012) Insertional mutagenesis by a hybrid piggyBac and sleeping beauty transposon in the rat. Genetics 192:1235–1248
Gilman JG (1972) Hemoglobin beta chain structural variation in mice: evolutionary and functional implications. Science 178:873–874
Gondo Y (2008) Trends in large-scale mouse mutagenesis: from genetics to functional genomics. Nat Rev Genet 9:803–810
Gondo Y, Fukumura R, Murata T, Makino S (2010) ENU-based gene-driven mutagenesis in the mouse: a next-generation gene-targeting system. Exp Anim 59:537–548
Graur D (2003) Single-base mutation—in nature encyclopedia of the Human Genome Macmillan Publishers Ltd
Green EL, Roderick TH (1966) Radiation genetics. In: Green EL (ed) Biology of the laboratory mouse. Dover Publications, New York, pp 165–185
Hitotsumachi S, Carpenter DA, Russell WL (1985) Dose-repetition increases the mutagenic effectiveness of N-ethyl-N-nitrosourea in mouse spermatogonia. Proc Ntl Acd Sc USA 82:6619–6621
Hoebe K, Beutler B (2005) Unraveling innate immunity using large scale N-ethyl-N-nitrosourea mutagenesis. Tissue Antigens 65:395–401
Hrabe de Angelis MH, Flaswinkel H, Fuchs H, Rathkolb B, Soewarto D, Marschall S, Heffner S, Pargent W, Wuensch K, Jung M, Reis A, Richter T, Alessandrini F, Jakob T, Fuchs E, Kolb H, Kremmer E, Schaeble K, Rollinski B, Roscher A et al (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. Nat Genet 25:444–447
Ivics Z, Hackett PB, Plasterk RH, Izsvák Z (1997) Molecular reconstruction of sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91:501–510
Ivics Z, Izsvák Z, Chapman KM, Hamra FK (2011) Sleeping beauty transposon mutagenesis of the rat genome in spermatogonial stem cells. Methods 53:356–365
Johnson FM, Lewis SE (1981) Mutation-rate determinations based on electrophoretic analysis of laboratory mice. Mutat Res 82:125–135

Justice MJ, Bode VC (1986) Induction of new mutations in a mouse t-haplotype using ethylnitrosourea mutagenesis. Genet Res 47:187–192

Justice MJ, Bode VC (1990) ENU-induced allele of brachyury (Tkt1) exhibits a developmental lethal phenotype similar to the original brachyury (T) mutation. J Exp Zool 254:286–295

Justice MJ, Zheng B, Woychik RP, Bradley A (1997) Using targeted large deletions and high-efficiency N-ethyl-N-nitrosourea mutagenesis for functional analyses of the mammalian genome. Methods 13:423–436

Justice MJ, Noveroske JK, Weber JS, Zheng B, Bradley A (1999) Mouse ENU mutagenesis. Hum Mol Genet 8:1955–1963

Justice MJ, Carpenter DA, Favor J, Neuhauser-Klaus A, Hrabé de Angelis M, Soewarto D, Moser A, Cordes S, Miller D, Chapman V, Weber JS, Rinchik EM, Hunsicker PR, Russell WL, Bode VC (2000) Effects of ENU dosage on mouse strains. Mamm Genome 11:484–488

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellåker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assunção JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294

Keays DA, Clark TG, Flint J (2006) Estimating the number of coding mutations in genotypic- and phenotypic-driven N-ethyl-N-nitrosourea (ENU) screens. Mamm Genome 17:230–238

Kiernan AE, Erven A, Voegeling S, Peters J, Nolan P, Hunter J, Bacon Y, Steel KP, Brown SDM, Guénet JL (2002) ENU mutagenesis reveals a highly mutable locus on mouse chromosome 4 that affects ear morphogenesis. Mamm Genome 13:142–148

Kile BT, Hentges KE, Clark AT, Nakamura H, Salinger AP, Liu B, Box N, Stockton DW, Johnson RL, Behringer RR, Bradley A, Justice MJ (2003) Functional genetic analysis of mouse chromosome 11. Nature 425:81–86

Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germline single-base-pair substitution in human genes. Am J Hum Genet 63:474–488

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sv USA 99:803–808

Largaespada DA (2009) Transposon mutagenesis in mice. Meth Mol Biol 530:379–390

Lewis MA, Quint E, Glazier AM, Fuchs H, De Angelis MH, Langford C, van Dongen S, Abreu-Goodger C, Piipari M, Redshaw N, Dalmay T, Moreno-Pelayo MA, Enright AJ, Steel KP (2009) An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat Genet 41:614–618

Lewis SE (1991) The biochemical specific-locus test and a new multiple-endpoint mutation detection system: considerations for genetic risk assessment. Environ Mol Mut 18:303–306

Lewis SE, Barnett LB, Sadler BM, Shelby MD (1991) ENU mutagenesis in the mouse electrophoretic specific-locus test, 1. Dose-response relationship of electrophoretically-detected mutations arising from mouse spermatogonia treated with ethylnitrosourea. Mutat Res 249:311–315

Lewis SE, Barnett LB, Shelby MD (1992) ENU mutagenesis in the mouse electrophoretic specific locus test. 2. Mutational studies of mature oocytes. Mutat Res 296:129–133

Liu SM, Leibel RL, Chua SC Jr (1998) Partial duplication in the Leprdb-Pas mutation is a result of unequal crossing over. Mamm Genome 9:780–781

Lu B, Geurts AM, Poirier C, Petit DC, Harrison W, Overbeek PA, Bishop CE (2007) Generation of rat mutants using a coat color-tagged sleeping beauty transposon system. Mamm Genome 8:338–346

Marshall RR, Raj AS, Grant FJ, Heddle JA (1983) The use of two-dimensional electrophoresis to detect mutations induced in mouse spermatogonia by ethylnitrosourea. Can J Genet Cytol 25:457–466

Martin N, Jaubert J, Gounon P, Salido E, Haase G, Szatanik M, Guénet JL (2002) A missense mutation in Tbce causes progressive motor neuronopathy in mice. Nat Genet 32:443–447

Mashimo T, Ohmori I, Ouchida M, Ohno Y, Tsurumi T, Miki T, Wakamori M, Ishihara S, Yoshida T, Takizawa A, Kato M, Hirabayashi M, Sasa M, Mori Y, Serikawa T (2010) A missense mutation of the gene encoding voltage-dependent sodium channel (Nav1.1) confers susceptibility to febrile seizures in rats. J Neurosci 30:5744–5753

Massironi SM, Reis BL, Carneiro JG, Barbosa LB, Ariza CB, Santos GC, Guénet JL, Godard AL (2006) Inducing mutations in the mouse genome with the chemical mutagen ethylnitrosourea. Braz J Med Biol Res 39:1217–1226

McDonald JD (1995) Using high-efficiency mouse germline mutagenesis to investigate complex biological phenomena: genetic diseases, behavior, and development. Proc Soc Exp Biol Med 209:303–308

McDonald JD, Trischler M, Stoorvogel W, Ullrich O (1994) The PKU mouse project: its history, potential and implications. Acta Paediatr Suppl 407:122–123

Menalled LB, Chesselet MF (2002) Mouse models of Huntington's disease. Trends Pharmacol Sci 23:32–39

Michaud EJ, Culiat CT, Klebig ML, Barker PE, Cain KT, Carpenter DJ, Easter LL, Foster CM, Gardner AW, Guo ZY, Houser KJ, Hughes LA, Kerley MK, Liu Z, Olszewski RE, Pinn I, Shaw GD, Shinpock SG, Wymore AM, Rinchik EM, Johnson DK (2005) Efficient gene-driven germ-line point mutagenesis of C57BL/6 J mice. BMC Genom 6:164

Mülhardt C, Fischer M, Gass P, Simon-Chazottes D, Guénet JL, Kuhse J, Betz H, Becker CM (1994) The spastic mouse: aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. Neuron 13:1003–1015

Muller HJ (1927) Artificial transmutation of the gene. Science 66:84–87

Nadeau JH, Singer JB, Matin A, Lander ES (2000) Analysing complex genetic traits with chromosome substitution strains. Nat Genet 24:221–225

Nolan P, Peters J, Strivens M, Rogers D, Hagan J, Spurr N, Gray IC, Vizor L, Brooker D, Whitehill E, Washbourne R, Hough T, Greenaway S, Hewitt M, Liu X, McCormack S, Pickford K, Selley R, Wells C, Tymowska-Lalanne Z et al (2000) A systematic genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. Nat Genet 25:440–443

Noveroske JK, Weber JS, Justice MJ (2000) The mutagenic action of N-ethyl-N-nitrosourea in the mouse. Mamm Genome 11:478–483

Oakberg EF, Crosthwait CD (1983) The effect of ethyl-, methyl- and hydroxyethyl-nitrosourea on the mouse testis. Mutat Res 108:337–344

Perez CJ, Jaubert J, Guénet JL, Barnhart KF, Ross-Inta CM, Quintanilla VC, Aubin I, Brandon JL, Otto NW, DiGiovanni J, Gimenez-Conti I, Giulivi C, Kusewitt DF, Conti CJ, Benavides F (2010) Two hypomorphic alleles of mouse Ass1 as a new animal model of citrullinemia type I and other hyperammonemic syndromes. Am J Pathol 177:1958–1968

Perez CJ, Dumas A, Vallières L, Guénet JL, Benavides F (2013) Several classical mouse inbred strains, including DBA/2, NOD/Lt, FVB/N, and SJL/J, carry a putative loss-of-function allele of Gpr84. J Hered 104:565–571

Peters J, Ball ST, Andrews SJ (1986) The detection of gene mutations by electrophoresis, and their analysis. Prog Clin Biol Res 209B:367–374

Pillers DA, Weleber RG, Green DG, Rash SM, Dally GY, Howard PL, Powers MR, Hood DC, Chapman VM, Ray PN, Woodward WR (1999) Effects of dystrophin isoforms on signal transduction through neural retina: genotype-phenotype analysis of Duchenne muscular dystrophy mouse mutants. Mol Genet Metab 66:100–110

Pretsch W, Favor J, Lehmacher W, Neuhauser-Klaus A (1994) Estimates of the radiation-induced mutation frequencies to recessive visible, dominant cataract and enzyme-activity alleles in germ cells of AKR, BALB/c, DBA/2 and (102xC3H)F1 mice. Mutagenesis 9:289–294

Quint E, Steel KP (2003) Use of mouse genetics for studying inner ear development. Curr Top Dev Biol 57:45–83

Quwailid MM, Hugill A, Dear N, Vizor L, Wells S, Horner E, Fuller S, Weedon J, McMath H, Woodman P, Edwards D, Campbell D, Rodger S, Carey J, Roberts A, Glenister P, Lalanne Z, Parkinson N, Coghill EL, McKeone R, Cox S, Willan J, Greenfield A, Keays D, Brady S, Spurr N, Gray I, Hunter J, Brown SDM, Cox RD (2004). A gene-driven ENU-based approach to generating an allelic series in any gene. Mamm Genome 15:585–591

Rinchik EM, Carpenter DA (1999) N-ethyl-N-nitrosourea mutagenesis of a 6- to 11-cM subregion of the Fah-Hbb interval of mouse chromosome 7: Completed testing of 4557 gametes and deletion mapping and complementation analysis of 31 mutations. Genetics 152:373–383

Runck AM, Moriyama H, Storz JF (2009) Evolution of duplicated β-globin genes and the structural basis of hemoglobin isoform differentiation in Mus. Mol Biol Evol 11:2521–2532

Runkel F, Hintze M, Griesing S, Michels M, Blanck B, Fukami K, Guénet JL, Franz T (2012) Alopecia in a viable phospholipase C delta 1 and phospholipase C delta 3 double mutant. PLoS ONE 7(6):e39203

Russell LB, Russell WL (1996) Spontaneous mutations recovered as mosaics in the mouse specific-locus test. Proc Natl Acad Sci USA 93:13072–13077

Russell LD, Ettlin RA, SinhaHikim AP, Clegg ED (1990) Histological and histopathological evaluation of the testis. Cache River Press, Clearwater

Russell WL (1962) An augmenting effect of dose fractionation on radiation-induced mutation rate in mice. Proc. National Acad. Sc. USA 48:1724–1728

Russell WL (1963) The effect of radiation dose rate and fractionation on mutation in mice. In: Sobels F (ed) Repair from genetic radiation damage, vol 4. Pergamon Press, New York, p 205–217

Russell WL, Kelly EM, Hunsicker PR, Bangham JW, Maddux SC, Phipps EL (1979) Specific locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. Proc Ntl Acad Sc USA 76:5818–5819

Russell WL, Hunsicker PR, Carpenter DA, Cornett CV, Guinn GM (1982a) Effect of dose fractionation on the ethylnitrosourea induction of specific-locus mutations in mouse spermatogonia. Proc Ntal acad Sc USA 79:3592–3593

Russell WL, Hunsicker PR, Raymer GD, Steele MH, Stelzner KF, Thompson HM (1982b) Dose—response curve for ethylnitrosourea-induced specific-locus mutations in mouse spermatogonia. Proc Natl Acad Sc USA 79:3589–3591

Sakuraba Y, Sezutsu H, Takahasi KR, Tsuchihashi K, Ichikawa R, Fujimoto N, Kaneko S, Nakai Y, Uchiyama M, Goda N, Motoi R, Ikeda A, Karashima Y, Inoue M, Kaneda H, Masuya H, Minowa O, Noguchi H, Toyoda A, Sakaki Y, Wakana S, Noda T, Shiroishi T, Gondo Y (2005) Molecular characterization of ENU mouse mutagenesis and archives. Biochem Biophys Res Commun 336:609–616

Schlager G, Dickie MM (1966) Spontaneous mutation rates at five coat-color loci in mice. Science 151:205–206

Schlager G, Dickie MM (1967) Spontaneous mutation and mutation rates in the house mouse. Genetics 57:319–330

Schmezer P, Eckert C (1999) Induction of mutations in transgenic animal models: BigBlue and Muta Mouse. Int Agency Res Cancer-Res Publ 146:367–394

Shedlovsky A, Guénet JL, Johnson LL, Dove WF (1986) Induction of recessive lethal mutations in the T/t-H-2 region of the mouse genome by a point mutagen. Genet Res 47:135–142

Shedlovsky A, King TR, Dove WF (1988) Saturation germline mutagenesis of the murine t region including a lethal allele at the quaking locus. Proc Ntl Acad Sc USA 85:180–184

Simon-Chazottes D, Tutois S, Kuehn M, Evans M, Bourgade F, Cook S, Davisson MT, Guénet JL (2006) Mutations in the gene encoding the low-density lipoprotein receptor LRP4 cause abnormal limb development in the mouse. Genomics 87:673–677

Stoye JP, Fenner S, Greenoak GE, Moran C, Coffin JM (1988) Role of endogenous retroviruses as mutagens: the hairless mutation of mice. Cell 54:383–391

Takahasi KR, Sakuraba Y, Gondo Y (2007) Mutational pattern and frequency of induced nucleotide changes in mouse ENU mutagenesis. BMC Mol Biol 8:52

Takeda J, Keng VW, Horie K (2007) Germline mutagenesis mediated by Sleeping Beauty transposon system in mice. Genome Biol 8(Suppl 1):S14

Takeda J, Izsvák Z, Ivics Z (2008) Insertional mutagenesis of the mouse germline with sleeping beauty transposition. Meth Mol Biol 435:109–125

Van Zeeland AA, Mohn GR, Mullenders LH, Natarajan AT, Nivard M, Simons JW, Venema J, Vogel EW, Vrieling H, Zdzienicka MZ et al (1989) Relationship between DNA-adduct formation, DNA repair, mutation frequency and mutation spectra. Annali dell'Instituto superiore di sanita (Ann 1st Super Sanita) Istituto Superiore di Sanita (ISDIS) 2003 25:223–228

Vogel EW, Natarajan AT (1995) DNA damage and repair in somatic and germ cells in vivo. Mutat Res 330:183–208

Vogel F, Rohrborn G (1970) Chemical mutagenesis in mammals and man. Springer, New York, p 519

Wahnschaffe U, Bitsch A, Kielhorn J, Mangelsdorf I (2005a) Mutagenicity testing with transgenic mice. Part I: Comparison with the mouse bone marrow micronucleus test. J Carcinog 4:3

Wahnschaffe U, Bitsch A, Kielhorn J, Mangelsdorf I (2005b) Mutagenicity testing with transgenic mice. Part II: comparison with the mouse spot test. J Carcinog 4:4

Wilsbacher LD, Sangoram AM, Antoch MP, Takahashi JS (2000) The mouse clock locus: sequence and comparative analysis of 204 kb from mouse chromosome 5. Genome Res 10:1928–1940

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH, Nachman NW, Pialek J et al (2011) Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43:648–655

Youssoufian H, Antonarakis SE, Bell W, Griffin AM, Kazazian HH Jr (1988) Nonsense and missense mutations in hemophilia A: estimate of the relative mutation rate at CG dinucleotides. Am J Hum Genet 42:718–25

Zheng B, Sage M, Cai WW, Thompson DM, Tavsanli BC, Cheah YC, Bradley A (1999) Engineering a mouse balancer chromosome. Nat Genet 22:375–378

Chapter 8
Transgenesis and Genome Manipulations

8.1 Introduction

In the early 1980s, the expression of *transgenic animals* was proposed to define animals having *foreign DNA sequences stably and deliberately inserted into their genome through human intermediaries*. With time and the advent of new techniques, this concept has progressively evolved, and nowadays, it is probably more appropriate to consider that transgenic animals are animals *whose genetic characteristics have been altered using one of the techniques of genetic engineering*. Whatever the definition, transgenic animals belong to the category of genetically modified or genetically engineered organisms (GMOs).

Transgenic mice can be created by using a variety of experimental procedures depending upon the aim of the experiment. Among these procedures, the microinjection of foreign DNA fragments directly into one of the pronuclei of embryos at the one-cell stage has been, and still is, widely used. Another popular technology, which was developed almost concomitantly, makes use of pluripotent stem cell lines derived from mouse embryos [embryonic stem (ES) cells], which can be cultivated and manipulated in vitro just like somatic cells and subsequently inserted into a blastocyst to participate in the formation of the germline of a chimeric organism. Transgenic animals have also been created by lentiviral infection of early embryos, by transposable elements, and by a few other techniques such as those recently reported that make use of specially designed site-specific nucleases.

Transgenic mice are produced routinely in an ever-increasing number of laboratories. They are also made to order by several private companies. All these transgenic animals have been invaluable for answering biological questions related to gene function and regulation. They are instrumental in the analysis of tissue differentiation and ontogeny, for example, by allowing the tracking of cell lineages. Finally, they allow the development of refined animal models of human genetic diseases.

In the previous chapter we concluded that the discovery of the mutagen ethyl-nitrosourea (ENU) could be considered a milestone in the history of mouse genetics, essentially because it made possible the creation of a virtually

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_8

unlimited number of new mutant alleles. Similarly, the advent of transgenic technology has been a true revolution, eliciting unprecedented changes in mammalian genetics and related fields. This chapter focuses on the production and use of transgenic mice.

8.2 Transgenesis Resulting from Pronuclear Injection of Cloned DNAs

The stable insertion of foreign DNAs into the germ line through microinjection into the pronuclei of fertilized mouse eggs was reported in the early 1980s in simultaneously several laboratories using the same technique but with different DNA molecules (Brinster et al. 1981; Costantini and Lacy 1981; Gordon and Ruddle 1981; Harbers et al. 1981; Wagner et al. 1981a, b). It was not until 1982 that the first transgenic mouse with a clear phenotype was developed by Palmiter, Brinster, and colleagues: a "giant" mouse carrying (and overexpressing) a rat growth hormone gene (Palmiter et al. 1982). Since these first descriptions, the technique has been improved and a variety of protocols for the efficient generation of transgenic mice has been published. Among the most popular "cookbooks" dealing with the subject, we recommend those by Hogan et al. (1994), and more recently by Hammes and Schedl (2000), Jackson and Abbott (2000), Houdebine (2003), Nagy et al. (2003), and Koentgen et al. (2010). We also recommend visiting the webpage of the International Society for Transgenic Technologies (ISTT) at http://www.transtechsociety.org/.

8.2.1 The Basic Experimental Protocol

The production of transgenic mice is achieved by injection, generally with a sharpened glass micropipette, of a few picoliters of a DNA solution (concentration ~2 ng/μl) directly into one of the pronuclei, while the egg proper (the zygote) is held, by suction, to another glass micropipette (the holding pipette). In most instances, the foreign DNA is injected into the male pronucleus because it is a little bigger and closer to the egg membrane than the female pronucleus.[1] In skilled hands, around 10–20 % of the microinjected eggs develop to term into a transgenic animal. Identification of the transgenic status is achieved by PCR amplification of DNA samples prepared from the presumptive transgenic animals with specific primers, and confirmation is obtained by using Southern blotting (Fig. 8.1).

[1] For this reason, the technique is sometimes designated "*pronuclear transgenesis.*"

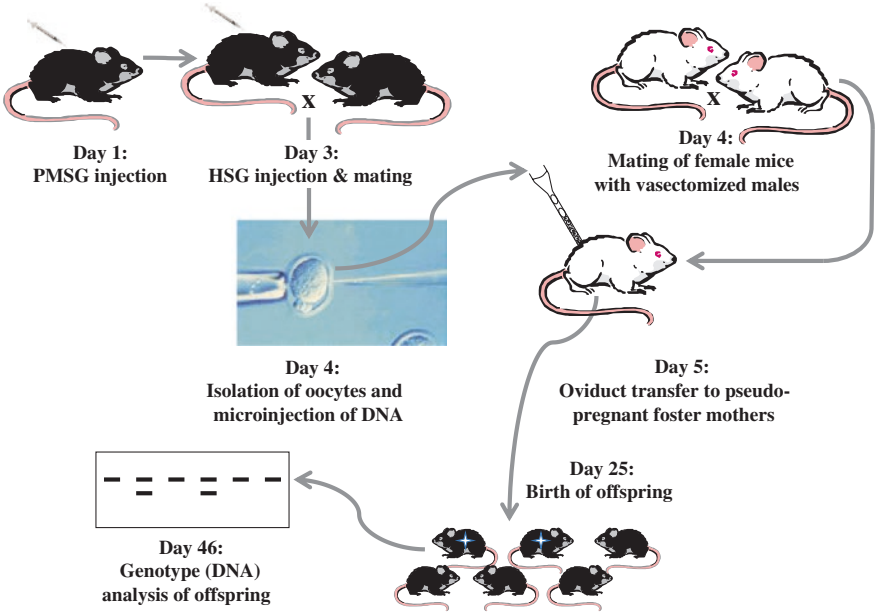


Fig. 8.1 *Producing transgenic mice by pronuclear injection*. The chart represents the different steps for the production of transgenic mice by pronuclear injection. Eggs are flushed out of the oviduct immediately after fertilization and then the transgene is microinjected in vitro with a glass micropipette. Once injected, the eggs are kept in vitro for a few hours and then transplanted into pseudo-pregnant females. Genotyping of the G0 (presumptive) transgenic mice can be achieved at any time from birth onwards. Every pup genotyped as positive by PCR (i.e., hemizygous Tg/0 carrier) should be considered a "founder," and independent lines should be developed from each founder

The DNA that is injected into the pronucleus can be either an unmodified or a natural copy of a gene cloned in its native genomic configuration, with its natural promoter, all its introns and other 5′or 3′ regulatory sequences, plus a few tenths of kb upstream and downstream of the sequences of interest. In most instances, however, the DNA that is used for transgenesis (the "transgene" proper) is artificial and designed in the laboratory according to the purpose of the experiment. It generally consists of several elements gathered in vitro, one piece at a time, then assembled using the most appropriate recombinant DNA technology. Finally, the transgene is cloned into a plasmid for amplification, mass production, and storage. When constructing such a fusion or chimeric gene for expression in transgenic mice, it is often easier to use a cDNA clone incorporating the coding sequences rather than the genomic DNA. This is especially true when the coding sequences in question stretch over a very long DNA segment or when they comprise many exons. Unfortunately, the levels of gene expression obtained with cDNA-based constructs are often lower than those obtained when genomic sequences are used.

Among the many explanations that can account for this observation, the existence of enhancers in the introns is the most likely (see Chap. 5).

Once selected, the relevant cDNA is placed under the control of a promoter, whose choice depends upon where and when it is desired that the transgene be expressed. When using cDNA (rather than genomic DNA) as a source of coding sequences, it is important to make sure that there is a translational start codon (AUG) within an upstream Kozak sequence (A/GCC**Pu**CC**AUG**G), which lies within the short 5′ untranslated region and directs translation of mRNA, and that there is an in-frame stop codon (UGA, UAG, UAA) for translational termination. Finally, it is also recommended to add an intron at the 5′ or 3′ end of the transgene because this allows the production of a more stable mRNA transcript and, finally, better transgenic expression (Brinster et al. 1988).

Experience teaches that the integration of the foreign DNA into the chromosome of the host probably occurs at random. In most instances, DNA integration occurs at the one-cell stage and at a single site but this is not a rule, and in 10–20 % of cases, the integration is delayed and occurs later during development. The mechanism of stable integration into the host genome is not precisely known, but it likely requires a double break (a nick) in the host (or recipient) DNA that is promptly repaired. Some scientists have suggested that this break might be the consequence of a trauma caused by the glass micropipette or by the injection of the DNA suspension. Even if this suggestion makes sense, it is probably not the only way for a transgene to integrate into a genome since delayed integrations, which are observed occasionally, are obviously not trauma dependent. When the foreign DNA does not integrate and stays isolated (as an *episome*, for example) in the nucleus for a few hours and integrates only at a later stage of development (2-cell; 4-cell), the organism develops as a mosaic. In this case, the detection of the transgene is more difficult and its transmission is unpredictable. In the case where the foreign DNA is present in all cells of the founder transgenic animal (noted F0, sometimes G0), it is then transmitted generation after generation as a new dominant "Mendelian" character.

The generic symbolic designation for a transgenic insertion is Tg. When the structure of the transgene is known, which is generally the case, a more precise designation applies. In this regard, we encourage the readers to refer to the guidelines for the standardized genetic nomenclature of transgenes in mice and rats at: http://www.informatics.jax.org/mgihome/nomen/gene.shtml#transg.

In contrast to gene and allele symbols, transgene symbols must not be italicized when they result from insertions of foreign DNA because they are not part of the native mouse genome.

The founder transgenic animals are hemizygous for the DNA segment (the symbol should be Tg/0, not Tg/–), and accordingly, the establishment of a "transgenic strain," in which the transgene is propagated by sexual reproduction, requires genotyping at each generation to avoid losing the transgenic DNA, unless the carriers have an obvious phenotype.

A method of safely maintaining a transgene in a mouse strain is to put it in the homozygous state, but this is difficult to achieve in practice. One reliable way

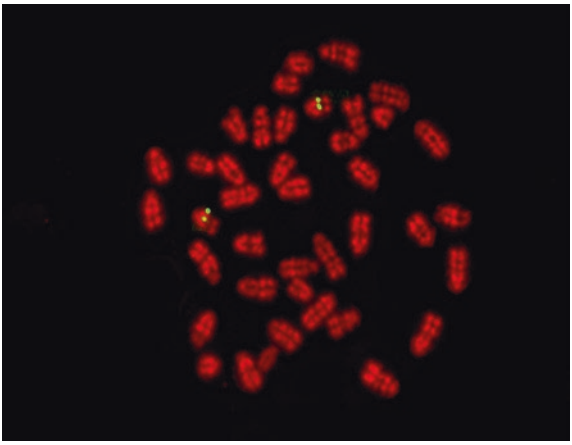


Fig. 8.2 *Fluorescence in situ hybridization*. Fluorescence in situ hybridization (*FISH*) with a transgene-specific probe indicates the localization of the transgene (*green dots*) in the karyotype (duplicated metaphase chromosomes). In this case, the transgenic insertion is homozygous (two copies). Using a chromosome or gene-specific probe with a different fluorescent staining allows for localization of the transgene on a specific chromosome (see Chap. 3)

of sorting out homozygous (Tg/Tg) from hemizygous (Tg/0) mice relies on the statistical analysis of their progeny when mated with a wild-type (WT or non-transgenic) partner (i.e., a progeny testing). A male mouse, identified as a carrier of the transgenic insertion based on a DNA test, producing only Tg/0 transgenic offspring in a progeny of 10 pups, when crossed with a non-transgenic partner has a greater than 90 % chance of being homozygous for the transgene (Tg/Tg). When the progeny size increases to 15, with only Tg/0 offspring, the probability increases to 99 %. Other possible means of identifying homozygous Tg/Tg mice are by quantitative real-time PCR (qRT-PCR) to determine zygosity and to distinguish hemizygous from homozygous transgenic mice (Ballester et al. 2004), or by cloning a segment of the DNA, flanking the transgene by inverse PCR and using it as a chromosomal marker for transgene localization. The transgenic insertion can also be visualized by in situ hybridization with a fluorescent dye (FISH) and accordingly located on a specific chromosome (see Chap. 3) (Fig. 8.2).

8.2.2 Factors Influencing Transgenic Expression

The number of copies of the transgene that integrates into the host genome is not controlled and ranges from one to several tens or even hundreds. Because sticky ends are generated when the foreign DNA is processed for injection, the cloned DNA copies are generally arranged in head-to-tail arrays in the transgenic insertion with frequent, and sometimes extensive, rearrangements generated in the

flanking regions. Using quantitative PCR technology, it is possible to roughly estimate the number of copies of the transgenic DNA; however, this is neither accurate nor reliable and, as we shall see, it sometimes changes with time.

As we already mentioned, investigators have no way of choosing the location where the foreign DNA will integrate. However, the integration site can seriously influence the transcription, and accordingly the expression, of the transgene. This is the case, for example, when the insertion site is in a heterochromatic or untranscribed (hypermethylated) region of the genome or when it is strongly influenced by a silencer sequence operating in its close vicinity. In these two cases, the transgene is weakly expressed or not expressed at all. Conversely, the sequences surrounding the transgene may contain regulatory elements acting on its promoter as enhancers of transcription. These enhancer sequences sometimes lead to an ectopic expression of the transgene; in other words, to an expression pattern that does not match with the spatial or temporal expression normally expected from the transcriptional regulatory elements the transgene contains.

These unexpected and somewhat erratic variations in expression are the consequences of a phenomenon known as the *position effect* and represent one of the main weaknesses of pronuclear transgenesis. The position effect and variations in copy numbers are two serious drawbacks, because both can affect transgenic expression. For this reason, in all cases, it is absolutely essential to make sure that the transgene is indeed fully expressed by checking whether all of the expected transcription products are present. One should also verify, as thoroughly as possible, that the structure of the transgene has not been affected by the mechanical handling of the DNA during the process of injection. This recommendation is especially important for large transgenes such as those made from yeast artificial chromosomes (YACs) or bacterial artificial chromosomes (BACs).

Since it is impossible to predict the effects of the integration site (i.e., the genomic environment) and of the number of copies on transgenic expression, it is highly recommended, when developing a transgenic strain for experimental purposes, to compare the offspring of several different founder transgenic mice and to consider only the features common to at least two independent strains as reliably attributable to the transgenic DNA. For this reason, it is not recommended to intercross mice originating from different founders but, on the contrary, to develop independent Tg lines from each founder.

Another classical observation when breeding transgenic animals is that 7–10 % of the transgenic insertions appear to be lethal when homozygous, presumably because a recessive lethal mutation (most probably a gene disruption) was mechanically generated at the time of integration in the recipient genome (insertional mutagenesis).

Finally, one must keep in mind that transgenic insertions are not always stable over time, and many investigators have reported the spontaneous and unexpected loss of the transgene from their favorite transgenic line. When a transgenic line is considered optimal and reliable, it is wise to preserve it as frozen sperm or embryos.

8.2.3 Using Transgenic Mice for Studying Gene Function and Regulation

A virtually unlimited number of transgenes can be engineered in vitro by the association of any coding sequence—normal or mutant—taken from any gene of any species, including plants and bacteria, and controlled by any regulatory elements. The use of transgenic mice is then a very convenient and efficient way to assess the function of genes. We will consider a few cases that have been selected as informative examples.

8.2.3.1 The Use of Transgenic Mice to Define the Function of Genes

Examples of this approach are provided by the homeogenes and the oncogenes, both of which are important actors in mammalian development. Homeobox-containing genes, the homeogenes, are transcriptional regulators with a remote ancestral origin, which are present in mammalian genomes and arranged in four paralogous clusters (*Hoxa*, *Hoxb*, *Hoxc,* and *Hoxd*). Because their structures are very similar, it was impossible to decide a priori whether each of these genes had a specific function, whether they had an effect because of the copy number (additive effect) or whether some of the copies were simple "backup" copies, preserved by evolution for unknown purposes. Transgenic mice were then made for some of these homeogenes with an intact coding sequence driven by a regulatory sequence different from the native one (driving ubiquitous expression, for example). In most instances, the embryos born with such extra transgenic insertions exhibited severe "homeotic" transformations indicating that indeed, most of the homeogenes in the *Hox* clusters had a specific function in the developmental patterning of the mouse embryo, a patterning reminiscent of their function in *Drosophila*, where they were initially discovered (Duboule 1998).

Transgenic mice have also been created with the coding sequence of (intact or mutated) oncogenes, or the sequence of genes whose function were not completely understood, downstream of a variety of regulatory sequences. Among these genes are the oncogenes *Abl1*, *Jun*, *Mos*, *Nras*, and *Myc*, as well as the tumor suppressor genes *Trp53* and *Rb*. Transgenic mice overexpressing oncogenes develop neoplasias in different tissues, depending on the promoter selected for the construct. For example, mice overexpressing the oncogene *Myc* driven by immunoglobulin enhancers develop lymphoid malignancies (Adams et al. 1985). The famous *OncoMouse*™ (the name is a trademark) is another example, but in this case, it carries the activated oncogene *v-Ha-ras* under the control of the MMTV promoter and, hence, produces mammary tumors (Hanahan et al. 2007). The subsequent analysis of these transgenic animals has provided an enormous amount of information concerning the role of these oncogenes in the regulation of several basic cellular functions and during the process of malignant transformation. The unique advantage of transgenesis in the case of homeogenes, oncogenes, and tumor

suppressor genes is to make the analysis of gene function(s) possible at the level of the whole organism.

8.2.3.2 Using Transgenic Mice to Identify and Characterize the Regulatory Sequences of Genes

While many mammalian genes are constantly and ubiquitously expressed, others are expressed in a tissue-specific manner, or only during embryonic life or only in the adult organism. Such variations in expression patterns occur because the genes are controlled by regulatory sequences that are in many cases, although not always, located in *cis* and upstream of the coding regions.[2] A good example of such tissue-specific regulation was reported for the gene encoding the cytokine leptin, which is expressed almost exclusively in adipocytes. After positional cloning of the mouse mutant gene *obese* (*Lep*^{ob}-Chr 6) (Zhang et al. 1994), it was demonstrated that the obese phenotype was a consequence of a nonsense mutation in codon 105 of the gene encoding the 16 kDa leptin protein. Researchers also learned that the highly tissue-specific expression of the *Lep* gene is controlled by a *cis*-acting regulatory sequence 161 bp long located upstream of exon 1 (He et al. 1995). For many genes, unfortunately, the regulatory sequences are not yet characterized and geneticists must design experiments to identify them accurately (see Chap. 5). This is important for a better understanding of gene regulation, of course, but it is also important if we consider that accumulating such data will certainly help in the future in silico identification of the regulatory elements based on sequence analogies.[3]

Transgenic mice are helpful for the identification of these regulatory sequences because experience teaches us that genes cloned in their native genomic configuration and introduced into the mouse germ line by transgenesis retain, in most instances, their tissue-specific and stage-specific patterns of expression, despite their integration at random sites. A popular strategy is to design in the laboratory a series of transgenes whose coding sequence encodes an easy-to-detect product which is not normally encoded in a mammalian genome (such a sequence is called a *reporter gene*), and to associate it by genetic engineering with a variety of regulatory DNA sequences, either upstream of the coding region, at the 5′ end or, less frequently, downstream of the 3′ end.

The gene encoding chloramphenicol acetyltransferase (CAT), from a transposon of *Escherichia coli*, has been extensively used to characterize the specific expression

[2] The genetic elements regulating gene expression are sometimes numerous and not always located in the close vicinity of structural genes. This explains (at least in part) why cloned structural genes, when used as transgenes, are sometimes regulated differently from the same genes in their natural, native environment (see Chap. 5). This point is inherent to transgenesis by in ovo injection and must always be kept in mind.

[3] In situ hybridization with labeled cDNAs is another way of analyzing the expression profile of a given gene.

associated with regulatory sequences because CAT activity can be assayed thanks to a very sensitive enzymatic test that has no background in eukaryotic cells (Overbeek et al. 1985). CAT has been progressively replaced by the gene encoding luciferase in the firefly (*Photinus pyralis*), largely because the assay to measure it is easier (Lira et al. 1990). *lac*Z, the historical gene encoding β-galactosidase of *Escherichia coli* (Goring et al. 1987), has been the cellular marker of choice to track cells in embryos and adults because of the ease of its detection and high cellular resolution in fixed embryos and tissues. The *lac*Z gene appeared to be particularly useful for studies of tissue- or position-specific gene expression. However, a major limitation is that *lac*Z cannot be used to mark cells in living tissues because the protocol to detect its expression requires tissue fixation. Fluorescent proteins offer advantages over enzyme-based reporters (e.g., *lac*Z, CAT) in the sense that their visualization does not require tissue fixation and is both quantitative and noninvasive. Indeed, fluorescent proteins make it possible to mark specific cells in living organisms, and also to follow such cells using fluorescence-imaging techniques (Fig. 8.3).

A classical reporter gene has been developed that consists of the sequence of the green fluorescent protein (GFP) of the jellyfish *Aequora victoria* (Misteli and Spector 1997). The product of this gene emits a green fluorescence elicited by direct illumination with blue light, and the analysis of the expression pattern requires neither fixation of the tissue nor cofactor or specific substrate, only UV light. Several variants of the wild-type GFP have been produced that emit in the blue (BFP), cyan (CFP), and yellow (YFP) regions. A series of variants derived from the red fluorescent protein (RFP) of the sea anemone *Discosoma sp.* are increasingly used because they emit a range of wavelengths in the red region, from the dark red of cherry to the yellow of banana. Interestingly, these different reporter genes can be combined allowing multiplexing and co-visualization

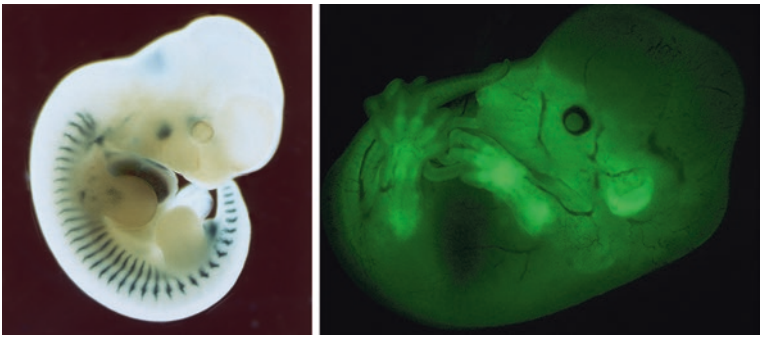


Fig. 8.3 *Analysis of gene expression with a reporter gene*. *Left* expression of the structural gene encoding *Lac*Z with regulation by the *Desmin* promoter. Observation of this embryo allows for detection of the tissues in which *Desmin*, a type III intermediate filament, is expressed (Courtesy C. Babinet). *Right* the embryo (recovered 13 days post-fertilization) is heterozygous for a knock-in allele in which the H2B-GFP coding sequence has been inserted in-frame into the gene encoding the platelet-derived growth factor receptor, alpha polypeptide ($Pdgfra^{+/H2B\text{-}GFP}$) (Courtesy J. Artus)

of several fluorescent proteins expressed in different tissues of a mouse (Passamaneck et al. 2006). Transgenic mice with reporter genes have been, and still are, extensively used by developmental geneticists (Lichtman et al. 2008). They have also greatly contributed to the annotation of the noncoding sequences.

8.2.4 The Use of Transgenic Technology to Generate Tissue- or Cell-Specific Ablations

Transgenic animals have been designed using tissue-specific regulatory sequences associated with sequences encoding cytotoxic proteins, with the aim of programming the genetic ablation of specific cell types either in the developing embryo or in the adult (Breitman et al. 1989). The most common strategy makes use of sequences encoding toxic proteins such as the A chains of the diphtheria toxin (DT-A) or of ricin (R-A), both of which block protein synthesis. In this case, the cytotoxic effect takes place as soon as the transgene is expressed. These studies indicate that programmed ablation of specific cell types can be stably transmitted generation after generation through the germ line (Breitman et al. 1987).

Another strategy has been developed that relies on the induced intracellular expression of the enzyme thymidine kinase (tk) of the herpes simplex virus (HSV). This enzyme is not directly toxic to animal cells but, unlike the mammalian thymidine kinase, it can phosphorylate certain nucleoside analogs such as acyclovir or ganciclovir, converting them into drugs that are toxic to dividing cells. In this particular case, the cell-killing effect becomes conditional since it depends both on the expression of the gene coding for viral thymidine kinase and on the administration of nucleoside analogs.

These methods of genetic ablation can be used to confirm the tissue specificity of a promoter; from this point of view, they appear complementary to the methods described above. Unfortunately, these methods, particularly the one using the highly toxic DT-A or R-A toxins as cell-killing agents, have a major drawback—the consequence of the extreme sensitivity of eukaryotic cells to these toxins. If the regulatory elements used in transgenic construction are not specific enough, a background expression of the transgene in cells that are not targeted results in misleading pathological conditions. This is mostly why, nowadays, this strategy of cell- or tissue-specific ablation has been abandoned for more specific approaches (see further).

8.2.5 Transgenic Complementation of a Mutant Allele Identified by Positional Cloning

As we already mentioned in the previous chapters, positional cloning of mouse mutations is an efficient approach for assessing the function of genes because the strategy directly associates a mutant phenotype with a specific gene. For example,

cloning a gene that is responsible for a leukodystrophy, once mutated, will point by definition to a gene involved in the development and organization of the white matter of the nervous system. However, when the candidate gene has only two alleles—one normal and one mutant—with the mutant being, for example, the consequence of a missense mutation (which occurs in about 75 % of cases), it is risky to conclude that the mutant allele is indeed responsible for the phenotype because there is always a chance, even if small, that the two observations (the phenotype and the mutation) are independent. In this case, it is generally necessary to prove that the missense allele is indeed causative of the pathology, and this can be achieved either by generating other alleles by mutagenesis (see Chap. 7 and later in this chapter) or by attempting to rescue the mutant phenotype by transgenic complementation. In this case, an appropriate breeding protocol is used to obtain genotypes that are certainly homozygous for the recessive mutation in question (*mut/mut*), normally leading to the deleterious phenotype, plus an additional (normal), functional transgenic copy of the candidate gene. The observation of a normal or nearly normal phenotype for this genotype validates the candidacy of the gene cloned by a positional approach. An example of transgenic rescue was reported endorsing the suspicion that a missense mutation in the gene encoding tubulin-specific chaperone E (*Tbce*^{pmn}-Chr 13) was indeed responsible for the deleterious phenotype of the mouse mutation *progressive motor neuronopathy* (Martin et al. 2002).

8.2.6 Using Transgenic Mice for Modeling Human Diseases

Different types of transgenic mice have been designed either to allow scientists to conduct experiments that were not possible with normal mice or to model a pathological condition that exists only in humans. We will provide a few examples to demonstrate the versatility of this transgenic technology.

8.2.6.1 Making Transgenic Mice Susceptible to Human Infectious Diseases

Poliovirus, the causative agent of poliomyelitis, infects primates but cannot spontaneously infect mice except for some type 2 virulent strains. Transgenic animals susceptible to all three poliovirus serotypes have been produced by pronuclear injection of the cloned human gene encoding the cellular receptor for the virus (Koike et al. 1991). These transgenic mice, when inoculated with poliovirus, mimic some of the clinical symptoms observed in humans and monkeys and are good models for studying the molecular mechanisms of pathogenesis of the virus as well as for testing vaccines against poliovirus infections.

Another example is the bacteria *Listeria monocytogenes*. These bacteria, once ingested by humans, can produce severe and sometimes fatal infections. The mechanisms by which the bacteria passes through the human intestinal barrier

is well known: It requires the intervention of a surface protein called *internalin*, which interacts with a host receptor, E-cadherin, to promote entry into intestinal epithelial cells. Murine E-cadherin, in contrast to human or guinea pig E-cadherins, does not interact with internalin, excluding the mouse as a model for experimental oral infection with *L. monocytogenes*. In contrast, in transgenic mice expressing human E-cadherin, internalin was found to mediate invasion of enterocytes and crossing of the intestinal barrier (Lecuit et al. 2001). These results illustrate well the value of transgenesis for understanding the physiopathology of human infections.

Models of the kind we have just described are, of course, of greatest interest for the study of infectious pathology, in particular for the development of efficient therapies and vaccines. Unfortunately, they illustrate rather exceptional situations and, in many cases where transgenesis was used to make animals susceptible to human pathogens, the situation has been discouraging. The determinism of susceptibility to infectious agents is sometimes complex and is rarely determined by the presence or absence of a single, species-specific cellular receptor. Progress in this area certainly awaits the discovery of genes whose products facilitate viral integration into the cell and full development of the replicative cycle. For example, scientists at the Rockefeller University and at the Scripps Research Institute demonstrated that the genes encoding CD81 and occludin were required for Hepatitis C virus (HCV) to enter human cells, and they demonstrated that making mice transgenic for these human genes made it possible to infect these transgenic animals with HCV (Dorner et al. 2011).

Even if perfect and faithful models cannot be made available simply by the mere addition of a few DNA segments, transgenic technology remains an interesting strategy to make progress in some aspects of infectious pathology. Transgenic technology, for example, allowed scientists to clarify the role of the complex cluster of genes encoding oligo-adenylate synthetase 1 (*Oas1*) in mouse susceptibility to flaviviruses (Scherbik et al. 2007; Simon-Chazottes et al. 2011) and has already provided insights into the pathogenesis of HIV-1.

8.2.6.2 Transgenic Models of Human Genetic Diseases

A mouse model of the human disease *osteogenesis imperfecta* type II (OMIM 166210) has been produced by injecting *in ovo* an abnormal mouse pro-α1 (I) collagen gene (*Col1a1*), orthologous to the abnormal human gene (Stacey et al. 1988; Pereira et al. 1993). The animals carrying such a transgene appeared very sick soon after birth, because of the modification of the extracellular matrix by the abnormal collagen fibers. In this case, the transgene had a dominant deleterious effect, the affected animals were almost impossible to breed, and the model proved to be of limited value. Nowadays, much better models can be generated using advanced techniques of transgenesis, as we will describe later.

A transgenic mouse strain has been created by pronuclear injection of both the normal human α-globin and the abnormal β^s-globin gene characteristic of sickle-cell anemia (Ryan et al. 1990). These animals were bred to β-thalassemic mice

to reduce endogenous mouse globin levels. When erythrocytes from these mice were deoxygenated, greater than 90 % of the cells displayed the same characteristic sickle shapes as erythrocytes from humans with sickle-cell disease. Compared to controls, the mice had decreased hematocrits, elevated reticulocyte counts, reduced hemoglobin concentrations, and splenomegaly, which are all indications of human sickle-cell disease. Such models are also of great help in the understanding of the pathophysiology of this debilitating disease as well as in the development of new drugs and therapies.

8.2.7 Transgenic Animals with Large DNA Inserts

Several techniques have been used to create mice transgenic for large DNA fragments. Among these techniques, the direct pronuclear microinjection of purified YACs or BACs has been the most popular (Jakobovits et al. 1993; Schedl et al. 1993; Lee and Jaenisch 1996; Van Keuren et al. 2009; Rossant et al. 2011). Such transgenic mice, when available, are very helpful for understanding the mechanisms operating when, for example, the genetic defect results from an unknown alteration occurring in a relatively large genetic region, or simply when the molecular origin of the defect is not completely clear. Several examples documenting the ability of wild-type alleles carried in YACs to complement mutations have been reported. The first one was the simple, complete rescue of the classical mouse albino mutation after injection into the germ line of albino (Tyr^{c}/Tyr^{c}) mice of a 250 kb YAC encompassing the wild-type mouse tyrosinase (*Tyr*) gene with all its introns and 155 kb of the 5′ flanking region (Schedl et al. 1992).

Original animal models of human genetic diseases have also been created using YAC transgenes. Among these, we must cite a model for Charcot–Marie–Tooth disease type 1A (Huxley et al. 1996) and a model for Huntington disease in which large intergenerational trinucleotide repeat expansions could be recreated, endorsing the use of these transgenic mouse models to refine the understanding of triplet repeat expansion and the resulting pathogenesis (Gomes-Pereira et al. 2011).

The possibility of inserting large-sized DNA fragments into the mouse genome will certainly be very useful for a better understanding of the phenotypic impact of the variations in genomic copy number (CNVs) (discussed in Chap. 5), as well as for the production of better models of Down syndrome (discussed in Chap. 3). Many fragments cloned from human chromosome 21 have been added to the mouse genome by *in ovo* transgenesis, producing phenotypes more or less reminiscent of those of human trisomy 21 (Smith et al. 1995; O'Doherty et al. 2005; Yu et al. 2010; Herault et al. 2012; Rueda et al. 2013). None of these models is perfect because of the complexity of the phenotype when several genes on different mouse chromosomes are used, but good progress is being made and transgenesis appears to be a technique of choice in this matter.

Many transgenic models of Alzheimer disease have been developed over the past several years. Most of these models replicate some of the pathological

features of the disease, such as plaque-like amyloid accumulations and astrocytic inflammation, but not all phenotypic aspects. In particular, the behavioral deficits are not faithfully modeled (Lithner et al. 2011).

Transgenesis with BACs or other large chromosomal segments is bound to become a very popular technology, with the foreseeable development of quantitative genetics in the years to come. The reason is that, unlike in the case of single Mendelian mutations, the genomic regions that have a quantitative effect on the phenotype are mostly unknown and, in this case, BACs containing the DNA segment where a quantitative trait locus (QTL) has been localized can be transferred into zygotes and the resulting mice tested for the quantitative trait in question. However, for this system to be applicable, BAC libraries must be available that contain the appropriate alternative alleles at the QTL in question (Heintz 2001; Abiola et al. 2003). Such libraries are now being prepared for different mouse species and strains.

8.2.8 Transgenic Knockdowns

In Chap. 5, when describing the different sorts of RNAs that are encoded in the mouse genome, we discussed the case of siRNAs and their possible use for gene silencing. Experiments of that kind have been undertaken several years ago by Katsuki et al. (1988) to assess the possibility of controlling gene expression by inducing the production of antisense RNAs in the genome. For their experiment, the Japanese scientists constructed a plasmid containing the promoter of the gene encoding the mouse myelin basic protein (MBP), followed by a portion of the rabbit β-globin gene associated with the mouse MBP-cDNA in the antisense orientation and a polyadenylation site. They observed that several transgenic mice for this transgenic construction had a phenotype similar to that of the mutant mouse *shiverer* (*Mbp*^{shi}-Chr 18). Antisense MBP messenger RNA was transcribed at high level in these mice, while the endogenous messenger RNA was reduced. The researchers concluded that the mice with an abnormal phenotype were *constitutive knockdowns* and that the transgene expression in vivo resulted in RNA interference (RNAi).

Since this first (successful) experiment, several other attempts at production of knockdown have been undertaken; some have been successful but most have failed. The reason is that, unlike in plants or invertebrates, double-stranded RNAs (dsRNAs) elicit an interferon response in mammals, resulting in global inhibition of protein synthesis and non-specific mRNA degradation. For this reason, short synthetic dsRNAs, whose length is below 30 bp, have been used to trigger the specific knockdown of mRNAs in mammalian cells without interferon induction. In the best experimental conditions, the efficiency of target knockdown can be as high as 90 % or greater, with permanent gene silencing in transgenic organisms indicating that the production of transgenic antisense RNA is an interesting approach to assessing gene function in vivo (Hitz et al. 2009).

8.2.9 Assessing the Mutagenic Activity of Chemicals with Transgenic Mice

As briefly mentioned in Chap. 7, at least two independent mice transgenic strains have been developed to assess in vivo the mutagenic activity of chemical substances of the environment: These strains are commercially available under the names of Big Blue® and Muta™Mouse (Wahnschaffe et al. 2005a, b). Transgenic mice of the Big Blue® strain have ~30–40 copies of the lambda LIZα shuttle phage vector integrated into their genome and the target for mutagenesis is the *lac*I gene. Muta™Mouse mice have ~80 copies of the lambda-gt10-*lac*Z shuttle vector integrated into their genome and the target is the entire *lac*Z gene.

Whatever the transgenic strain, the chemical compound to be tested is administered to the transgenic mice under several forms and at different doses. After a few days, DNA samples are then extracted from several tissues of the tested mice. The targeted genes are excised and packaged into lambda phage heads by using specially designed molecular kits, and the phages are transfected into bacteria. Finally, the transfected bacteria are plated on indicator plates containing selected chromogenic substances. Under these conditions, the phage-transfected bacteria with mutations in the targeted genes form plaques of a different color from those of bacteria with a non-mutated target gene, and the ratio of colored plaques to colorless plaques is a reliable measurement of the mutagenicity of the compound tested.[4] These transgenic strains have been extensively used and have provided fast and reliable estimations.

8.2.10 Mutations Induced by Pronuclear Transgenesis

As mentioned earlier, approximately 8–10 % of transgenic insertions result in recessive lethal mutations, and a much lower percentage in recessive viable. Good examples of the situation are two independent alleles at the *Formin* locus[5] (*Fmn1*-Chr 2) and a mutation described as *cryptorchidism with white spotting* (*crsp*-Chr 5) (Woychik et al. 1985; Messing et al. 1990; Overbeek et al. 2001). Such mutations by insertions would appear a priori to be interesting situations, considering that the inserted DNA (whose sequence is known) could be used as a tag for the identification of the mutated gene and accordingly for facilitating its positional cloning (for reviews, see (Jaenisch 1988; Gridley et al. 1990; Meisler 1992). In practice, however, the cloning of DNA flanking the insertion loci has often proved difficult as a consequence of the structural changes generated by the insertion.

---

[4] The phage-transfected bacteria with mutations in the *lac*I gene form blue plaques, whereas bacteria with a non-mutated *lac*I form colorless plaques in tests with the Big Blue® strain. With the Muta™Mouse strain, the basic principle is similar but the color of the plaques depends upon the experimental conditions.

[5] The first of the two alleles resulting from a transgenic insertion at the *Formin* locus (*Fmn*-Chr 5) has been known for a long time under the name of *limb deformity* (*ld*).

8.3 Generating Alterations in the Mouse Genome Using Embryonic Stem Cells

The technique of transgenesis by pronuclear injection of exogenous DNA sequences has been a true revolution in mammalian genetics. It has enabled hundreds of experiments that have provided answers to fundamental questions regarding the organization and functioning of the mammalian genome and has permitted the creation of many useful and original animal models for biomedical research. Unfortunately, the technique has some important limitations. One is that it allows the addition but not the deletion or substitution of genomic material meaning that, except in some rare situations, it is not possible to produce alterations with a recessive phenotypic expression. Another limitation is that the injected DNA inserts randomly in the genome of the host, and for this reason, the expression of the transgene often varies from one founder transgenic mouse to the next due to unique interactions with other genomic sequences in the background and disconnection of the transgene from its natural regulatory elements (see Chap. 5). Such limitations do not apply to the genetic alterations produced by using the techniques of genetic engineering in embryo-derived stem cells (ES cells). These techniques have been developed over the last 30 years and are still extensively used in the mouse for the production of a variety of targeted alterations. We will review the most commonly used.

8.3.1 Embryonic Stem Cells and their Advantages

ES cells were developed in the early 1980s (Evans and Kaufman 1981; Martin et al. 1981). They were derived from cells dissected from the inner cell mass (ICM) of blastocysts that were cultured in vitro, generally on feeder layers of fibroblasts, in tissue culture media supplemented with a few percent of fetal calf serum, with a high concentration of glucose, with glutamine and β-mercaptoethanol. To prevent these cells from differentiating in vitro, low concentrations of leukemia inhibitory factor (LIF) were added to the medium and the cells were re-plated at a relatively rapid pace.

ES cells represent a material of choice for geneticists because they can be manipulated (almost) like ordinary somatic cells, as long as they are maintained in vitro, while retaining all their developmental potentialities, in particular their capacity to differentiate into derivatives of all three embryonic germ layers (pluripotency). In addition, and most importantly, when merged with the cells of the ICM of a recipient blastocyst, many ES cells are capable of participating in the formation of chimeric embryos, and provided that these ES cells are euploid (i.e., with 2n chromosomes, a normal XY or XX complement, and no deletions or other types of chromosomal rearrangements), they are often capable of participating in the formation of the germ-cell lineage of the embryos in question. It is then

possible to apply to ES cells the classical techniques used in somatic cell genetics while they are in vitro (e.g., selection based on resistance or susceptibility to a specific drug), to isolate clones of cells with a pre-defined genetic characteristic, to "shuttle" them back into the germ line of a chimeric mouse, and finally to breed a strain of mice that have integrated into their genome an alteration engineered in vitro. The first experiments on genetic engineering with this type of cells were carried out by Gossler et al. (1986) and by Robertson et al. (1986). They were real breakthroughs,[6, 7] when these experiments were performed, most of the ES cell lines available for the purpose of scientific research were derived either from embryos of the 129/SvPas inbred strain (new nomenclature 129S2) or from the 129/J strain (new nomenclature 129P3/J). Nowadays, taking advantage of technological progress, especially in terms of culture conditions, many other ES cell lines have been derived from a variety of strains and most of them are stable and reliable, producing a high percentage of chimeric animals and a good germ line transmission ratio. The ES cell lines derived from strain C57BL/6N have become popular and have been selected in many transnational projects. This was a wise choice given that the reference sequence of the mouse genome is also from the C57BL/6 inbred strain.[8] ES cell lines derived from NOD, BALB/c, and some immunodeficient strains (such as NSG) are also available or under development. On the other hand, in the laboratory rat, the development of germ line-competent ES cells was only possible very recently (Ping et al. 2008).

Chimeras resulting from the fusion of an engineered ES cell with cells of the ICM of a recipient embryo can be identified, a few days after birth, for example, on the basis of their dappled coat color. This is very obvious when, for example,

---

[6] Well before the development of ES cells, another kind of cell, the *embryonal carcinoma* or EC cells, was used by oncologists and geneticists for investigating the genetics of cell–tissue differentiation. These cells were derived from spontaneous or experimentally induced testicular or ovarian teratocarcinomas (Stevens 1960). They were cultured in vitro, in the form of stable undifferentiated cell lines and then transplanted into mice of the same strain (syngeneic transplantation). Most of these cell lines, once engrafted, were able to differentiate into a variety of tissue (nervous tissue, bone, fat tissue, muscle, etc.), and some even proved able to participate in the formation of a chimeric organism (Papaioannou et al. 1975). They had, however, major drawbacks for the study of tissue differentiation: They were malignant and became rapidly aneuploid, and accordingly, they could not be used for the production of chimeric mice with germ line transmission.

[7] *Induced pluripotent stem cells* (iPSCs) are pluripotent cells derived from adult somatic cells after forced re-expression of some specific genes that are normally inactive. Such cells have been established in many species including human and mice. These iPSCs have many characteristics in common with ES cells and are being used in many experiments (for example, in the area of regenerative medicine). However, they have no obvious advantages over the long-established ES cells for the production of transgenic mice, and accordingly, they will not be considered in this chapter.

[8] The two strains C57BL/6N (ES cells) and C57BL/6J (genome sequence) are not completely identical, and recent estimates indicate a difference of ~1–2 % (SNPs) at the genome level (see Chap. 9).

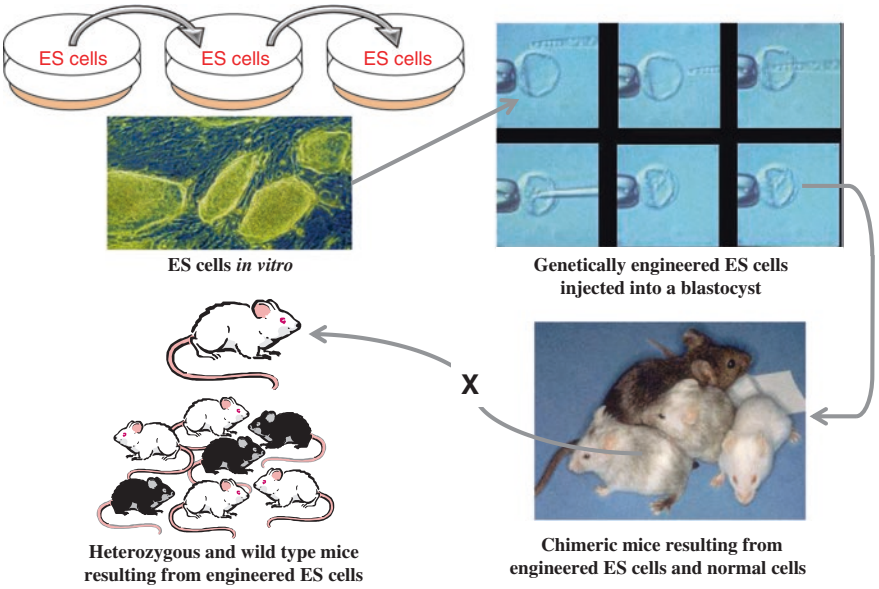


Fig. 8.4 *Targeted mutagenesis in the mouse using engineered ES cells*. The chart represents the different steps for the production of transgenic mice from genetically modified ES cells. ES cells can be cultured in vitro for several generations, remaining in an undifferentiated status. While in vitro, the ES cells can be manipulated like ordinary somatic cell lines and, in particular, can then be selected on the basis of specific criteria. ES cells can also be placed inside full-grown blastocysts where they spontaneously merge with the inner cell mass. Provided that the ES cells are still pluripotent and euploid, fertile chimeric mice can result from these reconstructed blastocysts. Mice with a dappled coat color in the figure are chimeras derived from blastocysts of (albino) hybrid mice (CSJF1) into which ES cells derived from a pigmented strain (129/Sv) were injected after several generations of in vitro culture. The size of the spots may vary according to the experimental conditions, but this does not faithfully reflect the percentage of chimerism in the germline. All of the other pigmented offspring of the chimeric mice are heterozygous for the genetic alteration(s) that may have been engineered in the ES cells. Two more generations are then necessary to observe the alteration in the homozygous state, and selection of the progenitors requires DNA genotyping

the ES cells are derived from the C57BL/6N inbred strain (which is non-agouti *a/a*—i.e., solid black) and the recipient blastocyst from either a wild-type (agouti *A/A*) or albino (Tyr^c/Tyr^c) strain. In these conditions, the chimeras exhibit a mixture of black and agouti (or albino) spots (Fig. 8.4).

Using coat color as a reference, one can estimate the percentage of chimerism, but a high level of chimerism does not necessarily correspond to a high rate of germ line transmission. Although chimeras can be from either sex, males are generally the only sex with germ line transmission because the majority of ES cell lines are XY. When grown in vitro for several generations, many (male) ES cells have a tendency to lose their Y chromosome and become XO.

8.3.2 Targeted Mutagenesis in ES Cells

The basic principle that characterizes targeted mutagenesis consists of applying a selection pressure on ES cells cultured in vitro that confers an advantage to the cells that may have lost (or acquired), spontaneously or after experimental manipulation, a characteristic encoded by a specific gene. The loss of a specific characteristic may result from a mutation, a deletion or any other kind of alteration, impairing the function of a given gene. The acquisition of a new heritable characteristic generally results from the transfection of foreign DNA molecules into the ES cells, followed by selection of the transfected cells based on a selective advantage conferred by the exogenous DNA. Once selected, the mutant or genetically modified ES cells are used for the production of chimeras, with the hope that a substantial proportion of the modified ES cells will still participate in the formation of the gametes. This will allow the production of transgenic mice with a targeted alteration in their genome.

8.3.2.1 The In Vitro Production of Mouse Models of Lesch–Nyhan Syndrome

Lesch–Nyhan syndrome (OMIM 308000) is a rare and severe X-linked metabolic disease in humans. The defect is characterized by the absence or inactivity of the enzyme hypoxanthine phosphoribosyl transferase (HPRT), an essential enzyme for the catabolism of purines. No animal model for this disease was available up to the mid-1980s, when two independent teams published the isolation of *Hprt*$^-$ clones of ES cells resulting from mutations in the *Hprt* gene. This was achieved after in vitro selection of *Hprt*$^-$/Y mutant cells that occurred spontaneously or after mutagenic treatment and became resistant to the toxic effect of the purine analog 6-thioguanine (6TG) when added to the culture medium.

Hooper et al. (1987) isolated a few *Hprt*$^-$ clones that occurred spontaneously and were selected with 6TG, injected them into blastocysts, bred chimeric mice and finally succeeded in establishing an *Hprt*$^-$ mutant strain. The isolation of clones of mutant ES cells by the mere in vitro selection on a particular phenotype (and genotype) was proved successful, although with a very low yield.

A technical improvement came from the use of mouse retroviruses as mutagenic agents and was a consequence of the early observations by Jaenisch and colleagues (1981). Two major conclusions of these pioneering experiments were that (i) retroviral vectors can be efficiently used as mutagenic agents for mammalian embryonic cells; and (ii) these vectors insert into the genome without generating extensive chromosomal rearrangements. Based on these observations, ES cells infected with the Moloney murine leukemia virus (M-MuLV) and mutant (null) alleles were recovered after selection with 6TG, at the same *Hprt* locus (*Hprt*$^-$), but this time, at a higher frequency (Kuehn et al. 1987).

The experiments reported above by Hooper and colleagues and Kuehn and colleagues were published simultaneously. They were the first experiments reporting the generation of a mutant strain in vitro, in ES cells, after selection of a particular phenotype. Surprisingly, however, the mutant mice, supposed to be a model of Lesch–Nyhan syndrome, did not exhibit any symptoms reminiscent of the human syndrome.[9, 10] From the genetic point of view, the result was somewhat disappointing but was nevertheless a great technical achievement, opening the way to many other technical refinements.

Considering the relatively high efficiency of the technique in terms of proviral integration numbers, massive infections of ES cells have been achieved from which embryos heterozygous for random insertions have been bred. These mutations by insertion have been put into the homozygous state using the classical two-generation micro-pedigrees (cross, backcross), and mutant phenotypes have been observed on some rare occasions. An interesting example is the recessive lethal mutation *Nodal*$^{tm1.1Mku}$ (Chr 10), with a block at the gastrula stage, which was found to be the consequence of a proviral insertion causing the loss of function of *Nodal*, a TGFβ-related gene (Lowe et al. 2001). Another mutation of the same kind (*Lrp4*^{dan}-Chr 2) was found to cause a syndrome of polysyndactyly as a consequence of the insertion of the proviral copy into the gene encoding MEGF7/LRP4, a member of the low-density lipoprotein receptor family (Simon-Chazottes et al. 2006) (Fig. 8.5).

The strategy that consists of infecting ES cells with M-MuLV, or any other kind of retrovirus, followed by the breeding of mice derived from the infected ES cells, allowed the identification of a few genes with effects on development. The retroviruses are mutagenic when they integrate into an exon or when they insert into an intron and disorganize the splicing process of the transcript encoded in the neighboring exons. An advantage in this case is that the retroviral insertion can also be used as a tag to identify DNA clones containing the mutated gene. Unfortunately, the yield of the strategy is low because, in most instances, retroviral insertions occur in noncoding regions and accordingly they have no direct or mechanical mutagenic effects. Another major drawback is that, for most autosomal genes in the mammalian genome, there is no efficient way to select in vitro the cells heterozygous for a recessive allele. In these conditions, it is necessary to breed mice homozygous for each proviral insertion and to unambiguously associate homozygosity for the proviral insertion with a specific phenotype, in general by the observation of tight linkage. This, however, is a tedious, risky and time-consuming enterprise.

---

[9] Mutations at the mouse *Hprt* locus probably occurred spontaneously in the past but were not recorded due to the complete absence of symptoms in the affected mice. We will never know for sure.

[10] The observation of differences (sometimes dramatic) in the symptomatology associated with a human syndrome and those observed in mice affected by mutations in the same orthologous gene is common. This, however, does not affect the value of the model.

Fig. 8.5 *Proviral insertional mutagenesis*. After experimental infection of ES cells with a defective Moloney retrovirus, a proviral copy was inserted, by chance, into the first intron of the gene encoding the low-density lipoprotein receptor related protein 4 (*Lrp4*-Chr 2). This insertional mutation disorganized the splicing process of the gene in question, making it virtually inactive. This resulted in the production of a fully penetrant, autosomal recessive mutation characterized by severe polysyndactyly (allele $Lrp4^{dan}$). The images on the *left* show normal paws from wild-type mice. On the *right*, the images depict paws from homozygous mutant mice with malformed digits and syndactyly

8.3.2.2 Another Model of Lesch–Nyhan Syndrome Resulting from Gene Targeting

In addition to the drawbacks mentioned above, one must also remember that one cannot target the integration of retroviruses at a specific site in the genome. In these conditions, the mutations generated are random and unpredictable. From this point of view, homologous recombination of extrinsic DNA molecules in ES cells resulting in the replacement of an endogenous gene by a different allele, in most cases non-functional, has been another breakthrough due to its potential applications. This technique is generally referred to as *gene targeting*.

The principle for the production of targeted mutations by homologous recombination is based on the observation that DNA fragments, once introduced into ES cells by an appropriate experimental procedure (e.g., electroporation or transfection), can recombine with the DNA of the host cells to become part of their genome. In most instances, the recombination occurs at non-homologous (or illegitimate) sites, but in some rare instances, it occurs at the homologous site. As a consequence, and provided that the transfected DNA molecules have been previously adequately modified by genetic engineering in vitro, a homologous recombination event can result in the replacement of an active and functional gene by an inactive one.

The idea that homologous recombination could occur in mammalian cells, and in particular in ES cells, originated from observations made in other eukaryotic organisms, in particular in the yeast *Saccharomyces cerevisiae*, where similar experiments had been successfully achieved. The detailed molecular mechanisms at work in the recombination process are not yet fully understood. It is likely that the mechanisms of homologous recombination overlap with those of illegitimate recombination, but a number of experiments indicate that they are not completely identical (for review, see Hooper 1992). Homologous recombination, of course, occurs at a much lower frequency than random integration (Smithies et al. 1985; Wong and Capecchi 1986). At this point, it should be noted that the idea of developing such a strategy was quite audacious if one compares the relatively small size of a cloned DNA that can be handled experimentally, to the gigantic dimensions of a mammalian genome!

To increase the yield of homologous recombination events, experience teaches us that the DNA molecule transfected into the ES cells must be linear, as large as technically possible, for instance up to 10 kb and more if possible, and should have the greatest possible length of sequence homology with the targeted DNA in the ES cell.

The first endogenous mouse gene that was modified by homologous recombination in ES cells was again the one encoding hypoxanthine-guanine phosphoribosyl transferase (*Hprt*-Chr X) (Thomas and Capecchi 1987). The experiment consisted of three steps. In the first step, a DNA molecule cloned from the *Hprt* targeted region and containing a few exons, the intervening introns and some flanking DNA sequences was cloned. In the second step, one exon in the cloned *Hprt*-DNA molecule was replaced by a piece of DNA of roughly the same size but with a different origin. Finally, the engineered cloned DNA was transfected into normal ES cells by electroporation. The idea underlying this manipulation was that, in the event of successful homologous recombination, the substitution of an exon by a segment of exogenous DNA would make the modified *Hprt* gene unable to transcribe a functional mRNA, thus generating a null allele.

While designing these "faked" or "counterfeit" DNA constructs to replace the targeted gene, scientists, instead of using segments of noncoding DNA as a foreign sequence, had the clever idea to use a minigene of bacterial origin encoding the enzyme neomycin phosphotransferase (neo^r) and capable of conferring to the transfected cells the capacity to resist to the toxic effect of neomycin. In these conditions, when plated in a culture medium with the antibiotic neomycin or, more precisely, with one of its amino glycoside analogs, G418, the normal ES cells were all killed while the cells synthesizing neomycin phosphotransferase (neo^r) resisted the cytotoxic effect of the drug. In other words, only those ES cells having stably integrated an engineered DNA molecule into their chromosomes, either at the targeted locus site or anywhere else in the genome, could survive. The rare ES cells clones where a strictly homologous recombination occurred would likely have reciprocally exchanged a functional copy of the *Hprt* gene for a non-functional one, and at the same time, they would also have acquired the property to resist the toxic effects of 6TG just like the $Hprt^-$ mutant cells reported above.

The advantages of this technique are twofold. The first is that, after selection with G418 (eliminating all cells with no stable DNA integration) and selection with 6TG (eliminating all cells with a functional *Hprt* gene), the only ES cells that would still grow in vitro are those where a homologous recombination event occurred. In other words, only the cells where the gene actually targeted has been effectively inactivated, or "knocked-out," would survive. The second advantage is that the mutation frequency by homologous recombination is higher than with any other technique. In the case reported above, for example, one stably transfected ES cell clone out of 150 was found to be a knockout (Capecchi 1989). This frequency of recombination events was considered high enough to adapt the technique to all cases where it was suitable for generating a null allele, even though the sorting out of the homologous recombinant ES cells from the non-homologous recombinant cells could not be achieved by the same, in vitro selection as in the case, we just reported for *Hprt*$^-$ cells.

Since these early experiments, thousands of genes have been inactivated using the gene-targeting strategy.[11] Genes inactivated by homologous recombination in ES cells are now collectively designated by the name of "*knockout*" or "*knock-out"* (KO). The in vitro engineered DNA molecule used for targeting the homologous native counterpart in the chromosome of the ES cells is designated the "*recombination vector*" Nowadays, in all experiments of this kind, confirmation that the expected event of homologous recombination actually occurred in the manipulated ES cells is sought by PCR amplification of critical DNA fragments with an appropriate set of primers followed by sequencing and confirmation by Southern blotting. The ES cells in question are then placed into a recipient blastocyst for the production of a chimera. The genetically engineered ES cells, once confirmed "reliable" and capable of participating in the germ line of the chimeric mouse, are stored deep-frozen for future use or distribution to the community.

8.3.2.3 Generating a Variety of Knockout Alleles by Homologous Recombination

Many of the knockout mutations that have been generated in mouse ES cells over the past several years have resulted from the use of replacement vectors as described above. In this case, after homologous recombination, the targeted gene is deleted by one of its specific coding sequences, which is replaced by a heterologous DNA that is, in many cases, a *selection cassette*. As a consequence of this substitution, the gene is inactivated and, at the same time, the manipulated ES cells acquire a selective advantage over a drug and can be positively selected. Several variations on this basic scenario have been used, and it is impossible to describe them all in this chapter. However, we can say that most of these strategies

[11] For their discoveries of the *principles for introducing specific gene modifications in mice by the use of embryonic stem cells* Drs. Mario Capecchi, Martin Evans, and Oliver Smithies were awarded the Nobel Prize in Medicine or Physiology in 2007.

consist of using a variety of selection cassettes, making use of bacterial genes encoding either resistance to hygromycin *B* or puromycin as alternatives to the *neo*^{r} cassette.

The design of the selection cassettes in the replacement vectors for homologous recombination depends on the nature of the targeted gene. If the gene in question is transcriptionally active in the ES cells, then the selection cassette is transcribed and positive selection with a drug can operate. However, if the gene in question is not expressed in the ES cells or if its expression pattern is unknown, it is then necessary to design a vector that incorporates a promoter active in ES cells, allowing the gene to be "switched on" when requested.

Replacement vectors allowing positive/negative selection have also been designed by inserting a *neo*^{r} mini-gene between two regions of homology, and inserting a gene encoding herpes simplex virus thymidine kinase (HSV*tk*) outside of the regions of homology. ES cells that are transfected in vitro with such a replacement vector are then subjected to a double selection: (i) the first one with G418, inducing the destruction of all ES cells that had not integrated at least one copy of the vector (and accordingly the *neo*^{r} mini-gene); and (ii) the second selection with the guanosine analog ganciclovir (GANC, sometimes spelled gancyclovir), killing the cells containing a functional thymidine kinase (*tk*) gene. This second level of selection eliminates the ES cells, in which non-homologous recombination occurred because in this case the HSV*tk* component of the vector is generally retained, while it is deleted after homologous recombination. In these conditions, only the very few cells where homologous recombination occurred can survive (Figs. 8.6 and 8.7).

The techniques for gene inactivation which we just mentioned have been described in detail in several review papers and book chapters (Hooper 1992; Hasty et al. 2000; Babinet and Cohen-Tannoudji 2001; DeChiara 2001; Koentgen et al. 2010). The use of replacement vectors with a selectable marker has been and still is very popular for the generation of null alleles because it is relatively

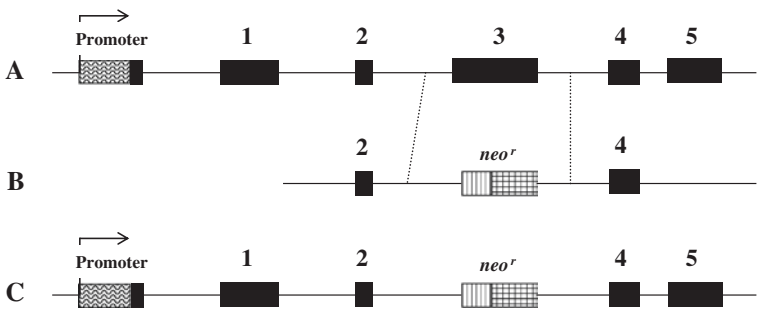


Fig. 8.6 *Gene targeting with a replacement vector 1*. Recombination events occurring in the regions flanking the *neo*^{r} cassette result in the deletion of exon 3 and its replacement by the *neo*^{r} cassette. The *neo*^{r} cassette confers a selective advantage on the recombined ES cells. The longer the sequence homology between the replacement vector and the host DNA, the better

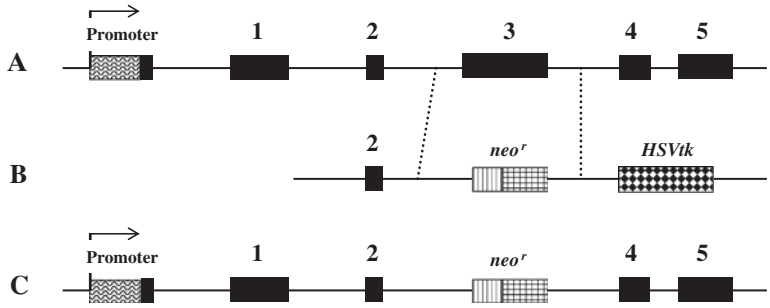


Fig. 8.7 *Gene targeting with a replacement vector 2*. Gene targeting with a replacement vector engineered with a positive/negative selection cassette. After homologous recombination, the *HSVtk* cassette is deleted while the *neo*^{r} cassette replaces exon 3. This recombination confers to the recombinant ES cells a selective advantage to *G418* and a selective disadvantage to *Ganciclovir*

straightforward and produces stable, permanent alterations. Unfortunately, cases have been reported where the selection cassette alters to some extent the expression of neighboring genes.

8.3.2.4 Generating Point Mutations by Homologous Recombination in ES Cells

The strategies described above, which make use of replacement vectors, require the introduction of extrinsic DNA sequences of various sizes into the genome of ES cells. Although mostly unknown, the consequences of this manipulation may have some possible adverse effects. This is why scientists have developed an alternative strategy, in two steps, leading to the creation of specific base-pair changes (missense or nonsense) in a specific DNA sequence, allowing the generation of so-called *knock-in* (KI) animals.[12]

The strategy in question is based on two successive steps of homologous recombination, with positive and negative selection, and makes use of mutant *Hprt*$^-$ ES cells similar to those resulting from the experiments reported above (Hooper et al. 1987; Kuehn et al. 1987) and two replacement vectors. The first replacement vector is designed to replace an exon of the targeted gene in HPRT-deficient (*Hprt*$^-$) cells with a functional *Hprt* minigene after the first homologous recombination (Selfridge et al. 1992).[13] After this first replacement, the recombinant ES cells are no longer resistant to the toxic effect of 6-thioguanine (6TG) and can grow

[12] The definition of *knock-in* also applies to the targeted insertion (and substitution) of any coding sequence at a particular locus of an organism. In these conditions, and in most instances, the inserted coding sequence is controlled by the regulatory regions of the targeted gene.

[13] The HPRT mini-gene is a selection cassette that is unique, since selection may be applied for its presence or absence.

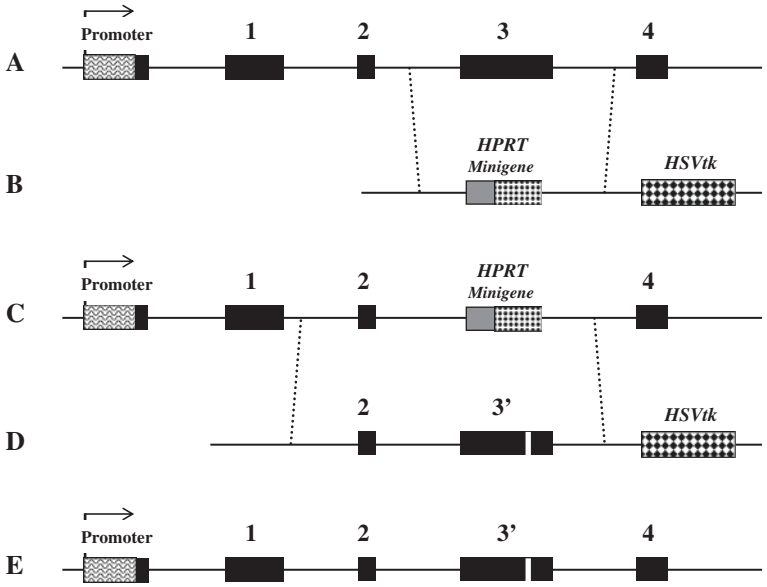


Fig. 8.8 *Induction of point mutations*. Induction of point mutations with two replacement vectors in *Hprt*$^-$ mutant ES cells. The first replacement vector substitutes an *Hprt* (functional) mini-gene for exon 3 and confers resistance to HAT (hypoxanthine, aminopterine, thymidine). The second recombination replaces the *Hprt* mini-gene by a mutated exon 3 (exon 3′) engineered in vitro. The ES cell then becomes sensitive to HAT but insensitive to 6-thioguanine. This homologous recombination is a *knock-in* (*KI*) because the original gene is replaced by a modified version, even if the gene is merely a mutant allele with only a point mutation

normally in so-called Littlefield's hypoxanthine, aminopterin, and thymidine (HAT) culture medium because they have a functional HPRT.[14] In these recombinant *Hprt*$^+$ cells, the targeted gene is deleted by one exon after replacement by the *Hprt* mini-gene. A second replacement vector is then designed that corresponds perfectly to the sequence of the original targeted gene, with the exception of a single base pair difference (an SNP) in the targeted exon.[15] This vector is synthesized in vitro, using a PCR technique of directed mutagenesis that is now routine in most laboratories. After this second replacement, homologously recombined ES cells are killed in HAT medium but survive selection by 6-thioguanine (6TG), as in the case of the original *Hprt*$^-$ deficient ES cells (Stacey et al. 1994) (Fig. 8.8).

Finally, and although it has no deleterious effects in the mouse, if the *Hprt*$^-$ mutation is considered undesirable, it can be easily eliminated by two rounds of sexual reproduction once the "offspring" of the mutated ES cells are born.

[14] *Hprt*$^-$ cells cannot grow in HAT medium because aminopterin blocks the endogenous synthesis of both purines and pyrimidines.

[15] Mice of this type are not transgenic animals sensu stricto because they do not have any exogenous DNA sequences "stably inserted into their genome." However, they are still GMOs.

This sophisticated technique of double replacement has potential applications, among which is the generation of a series of co-isogenic strains of mice (see Chap. 9). This can be achieved by using the same *Hprt*$^-$ ES cells, which are derived from a highly inbred strain, then by targeting these cells with a variety of second-set vectors from different inbred strains (replacing the *Hprt* mini-gene). One can then generate a variety of different point mutations in the same genetic background (the same inbred strains).

Alternative techniques using an insertion vector instead of a replacement vector, and known as the *hit-and-run* or *in-and-out* techniques, have been used for the generation of point mutations in targeted genes (Hasty et al. 1991; Valancius and Smithies 1991). These techniques required two rare intra-chromosomal recombination events to occur, and accordingly, they appeared to be less efficient than the technique making use of two replacement vectors. For this reason, they have been abandoned.

8.3.2.5 Knock-Ins Are Sometimes Sophisticated Knockouts

An interesting variation of the technique used for obtaining knockout mutations has been designed by introducing, via the replacement vector, the coding sequence of reporter genes in-frame with the promoter of the targeted gene. To give an example of the high degree of sophistication of this method, we refer to an experiment designed to assess the function of the genes encoding *connexins* (Filippov et al. 2003). Connexins are expressed in the various cell types of the central nervous system and are thought to regulate some of the functional properties exhibited by immature and mature cells. Understanding the specific role of each connexin in these processes required an unambiguous characterization of their spatial and temporal pattern of expression. To achieve this aim with connexin 26 (CX26) (gene symbol *Gjb2*, for gap junction membrane channel protein beta 2), scientists generated a reporter allele ($Gjb2^{lacZ}$) in which the pattern of expression of the gene encoding β-galactosidase was controlled by the endogenous *Gjb2* promoter. Then, by observing $+/Gjb2^{lacZ}$ heterozygous mice, the researchers could easily identify the tissues expressing CX26 (i.e., liver, kidney, skin, cochlea, small intestine, placenta, and thyroid gland) and demonstrated that the expression of CX26/*Gjb2* is restricted to the meninges both in embryonic and adult brain. The same researchers also noted that homozygous $Gjb2^{lacZ}/Gjb2^{lacZ}$ knockout embryos died early in utero, indicating that at least one intact copy of the *Gjb2* gene is necessary for normal embryonic development.

Such a mutation, where a gene is inactivated by the insertion of a foreign coding sequence driven by the same promoter, is also designated as a *knock-in*.[16] The knock-in strategy is universal and can be applied to any gene to inactivate it and, at

[16] In short, the main difference between a *knock-out* and a *knock-in* allele is that, in the case of a *knock-in,* the gene product is different from the normal allele but still has a function, even if the function in question is totally unrelated to the function of the original allele. In the case of a *knock-out*, the gene has simply been made inoperative.

the same time, visualize its expression pattern in the developing embryo or in the adult. The knocked-in genes are in general more faithfully expressed than the transgenes produced by pronuclear injection.

8.3.2.6 Engineering Conditional Knockout Mutations—The Cre-*lox*P Strategy

When produced by using one of the techniques described above, knockout mutations affect all the cells of the developing embryo in which the gene is normally expressed, starting from the early stages of development. For this reason, the mutations in question are often designated *constitutive knockouts*. Since most of the knockout alleles behave as recessives, the situation is in general well tolerated as long as the allele stays heterozygous. However, when the knockout allele is homozygous, the gene is permanently switched off in all cells, and the situation may become problematic. This is the case, for example, when the knockout allele results in early embryonic lethality because this hinders the analysis of the gene function(s) in later developmental stages or in the adult. It is also a drawback when the inactivation of the targeted gene results in the deregulation or misregulation of the expression of other genes.

To bypass these drawbacks, gene-targeting strategies have been developed that allow the (knockout) mutations to be made conditional (*conditional knockout* or *cko* mice). With conditional mutations, both the timing of gene inactivation and the cells or tissues in which the gene is to be "switched off" can be controlled. The discovery and development of these techniques has been another fundamental achievement in transgenesis.

The strategies used for the production of conditional knockouts make use of two transgenic strains: one in which the targeted gene is modified in a way that ensures its future inactivation and the other where the time- or tissue-specific expression of the mutation is programmed. Each of the two strains is normal and fully viable, but when intercrossed, all the ingredients necessary for inactivation are merged into the genome of their offspring.

The most popular strategy is known as the Cre-*lox*P strategy and makes use of Cre recombinase (from cyclization recombinase), a 38 kDa enzyme derived from the bacteriophage P1 (Utomo et al. 1999; Nagy 2000). Cre recombinase cuts and recombines the DNA strand at specific sites called *lox*P sites (short for locus of X-ing over P1) (Sauer 1993). These *lox*P sites consist of two 13 bp inverted (palindromic) repeats separated by an 8-bp asymmetric spacer region that defines the orientation of the site. Such sites do not exist in the mammalian genome (Fig. 8.9). When the *lox*P sites are in the same orientation and on the same strand (or chromosome), the intervening stretch of DNA is excised as a circular loop. When two *lox*P sites are in opposite orientations and on the same chromosome, the intervening DNA segment is inverted. Finally, when the *lox*P sites are on two different chromosomes, the recombinase generates a reciprocal translocation. When there are more than two *lox*P sites in the same genome, a variety of recombinations can occur.

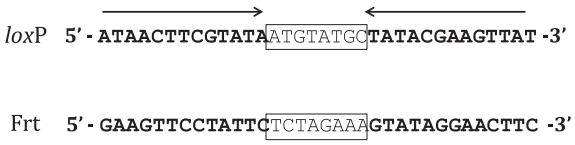


Fig. 8.9 *loxP and Frt sites*. A *lox*P site (*top*) consists of two 13-bp palindromic sequences (*arrowed*) flanking an 8-bp spacer region (*boxed*). These 8-bp define the directionality of the *lox*P site. When two *lox*P sites are placed on the same strand and in the same orientation, the Cre recombinase deletes the intervening sequence plus one *lox*P site. When the sites are in opposite orientations, Cre generates an inversion of the intervening sequence and both *lox*P sites are retained. When the *lox*P sites are on different chromosomes, the Cre-recombinase generates a reciprocal translocation. Nucleotide sequence of the 34-bp-long *FRT* site (*below*). The palindromic sequences bind the recombinase, whereas the spacer is the site of DNA break, exchange, and ligation

To illustrate the basic principle of the method, we will take a historical example: the case of T-lymphocyte-specific inactivation of the gene encoding the DNA-directed βpolymerase (*Polb*-Chr 8)(Gu et al. 1994). In this experiment, a strain of mice (strain A) had its *Polb* gene specifically modified by targeted homologous recombination with a replacement vector. The replacement vector was designed in such a way that an essential sequence of the *Polb* gene, actually the promoter and the first exon, became flanked by two *lox*P sites. The replacement vector was also designed in such a way that it contained two selection cassettes: a *neo*^{r} cassette and a thymidine kinase (HSV*tk*) cassette, themselves flanked by a third *lox*P site as indicated in Fig. 8.10. After homologous recombination, the targeted gene, *Polb,* ended up with three *lox*P sites inserted in the same orientation: the first one upstream of the promoter and exon 1, a second one in intron 1 upstream of the selection cassettes, and a third site downstream of the cassettes but upstream of exon 2. As geneticists say, the gene was then *floxed* (flanked by *lox*P sites) but, at this point, it was still functional and normally transcribed, and the mutation was only cryptic, or "premeditated", so to speak. The *neo*^{r} and HSV*tk* cassettes were useful for positive/negative selection with the classical drugs G418 or ganciclovir, should it be necessary.

Concurrently, another strain of mice (strain B) transgenic for a gene encoding Cre-recombinase was produced by classical pronuclear microinjection. The Cre-encoding transgene in this case was driven by a lymphocyte creatine kinase (*lck*) promoter, which is specific for T cells. When strains A and B were intercrossed, generating double transgenic (*bigenic*) mice, the product of the Lck-Cre transgene triggered deletion of the floxed segment in one or both chromosomes according to the genetic constitution (heterozygous or homozygous) of strain A, but in T cells exclusively. The consequences of the mutation (symbolized *Polb*$^-$)[17] on T cells could then be analyzed because mutant mice were viable, whereas they would have died if the mutation had been expressed ubiquitously during development.

[17] According to the official nomenclature rules, the symbol for this mutation should be *Polb*$^{tm1.1Rsky}$. This was the first targeted mutagenesis at this locus in Rajewsky's laboratory.

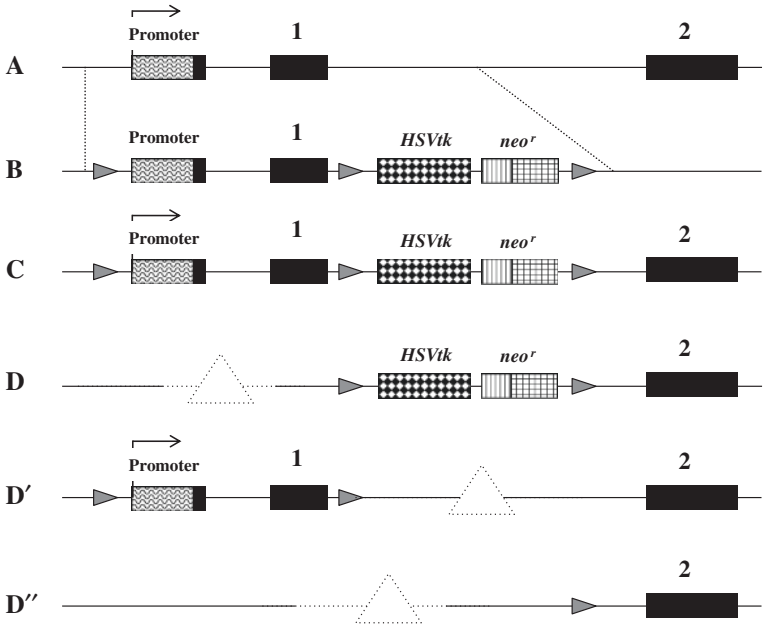


Fig. 8.10 *Inducing gene-targeted deletions with the Cre-loxP system*. In this experiment, the replacement vector (*B*) was designed in such a way that the *Polb* targeted region ended up with three *lox*P sites inserted in the same orientation: the first one upstream of the promoter and exon 1, a second one in intron 1, upstream of the selection cassettes, and a third one downstream of the cassettes but upstream of exon 2 (*C*). When Cre is synthesized, the segments flanked by two *loxP* sites (the *floxed* regions) are deleted, producing three different types of ES cells (*D*, *D′*, *D″*). The ES cells in which the targeted gene is deleted (and permanently inactivated—*D* & *D″*) are the most interesting. The *neo*^{r} and HSV*tk* cassettes were useful for positive/negative selection with the classical drugs G418 and *Ganciclovir*

Hundreds of experiments of the type described above, leading to tissue- or cell-specific gene inactivation, have been performed in recent years using either the Cre-*lox*P system or a similar system known as FLP-*Frt* (FLP for Flippase recombination enzyme—*Frt* for Flippase Recognition target). The FLP-*Frt* system is very similar to the Cre-*lox*P system but makes use of a yeast recombinase with another specific restriction site.

With these systems, an unlimited number of mutations may be designed, keeping in mind that Cre (or FLP) deletes any DNA segment once the latter is flanked by two *lox*P (or *Frt*) sites, provided these sites are oriented in the same direction. When there are more than two *lox*P sites in the same cell, as is the case in Fig. 8.10, Cre cuts at each site and, under specific conditions, generates a variety of deletions or translocations. Selection can then be applied to retain one cell type and not the others if selection cassettes have been judiciously inserted in critical regions (Gu et al. 1994).[18]

[18] This explains why, with such molecular tools, any kind of chromosomal rearrangement can be engineered in vitro. In the past, these chromosomal rearrangements were occasionally collected in the progenies of mice after irradiation in the post-meiotic stages (see Chap. 3).

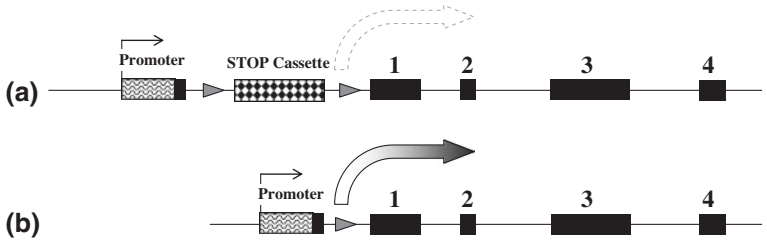


Fig. 8.11 *Cre-loxP regulation of transcription*. **a** A floxed "stop" cassette hampers transcription of the gene downstream. **b** When the "stop" sequence is deleted by the action of the Cre-recombinase, transcription resumes

A similar strategy has been employed using the same strain A (with floxed *Polb*) and another strain (strain C) with the interferon-inducible promoter of the gene *Mx1* to regulate Cre expression. After crossing strain A with strain C, *Polb* inactivation was induced in adult animals after interferon treatment. In this case, inactivation was complete in liver, spleen, and bone marrow while it was incomplete in other tissues (Kuhn et al. 1995). These experimental results demonstrated that Cre-mediated recombination could also be effectively induced in nondividing cells. The expression of the Cre transgene can be made inducible, adding more sophistication to the system. The tamoxifen-inducible Cre^{ERT2}, which can be activated by administration of tamoxifen to the transgenic mice, is very popular (Feil et al. 2009). Nowadays, many Cre-expressing lines are being produced as knock-in mice that incorporate the Cre sequence into the gene of interest (instead of creating transgenic lines using pronuclear microinjection).

The Cre-*lox*P strategy can also be used to regulate the expression of a specific protein in a tissue- or cell-specific way using a strategy that is schematically outlined in Fig. 8.11. In this example, the *lac*Z gene is a reporter gene driven by a ubiquitous promoter (e.g., *Rosa 26*) with a floxed "stop" sequence inserted between the promoter and the *lac*Z coding sequence. The "stop" sequence is a short segment of DNA with several terminator codons that impede translation of the protein. When the floxed "stop" sequence is deleted by the action of Cre in some specific cells or tissues, then the *lac*Z gene is transcribed following the same pattern of cell/tissue specificity (Lakso et al. 1992; Pichel et al. 1993) (Fig. 8.12).

To add versatility to the method, it must be kept in mind that both the Cre and FLP recombinases can be used, simultaneously or successively, in the same experiment.

Since experiments on conditional targeting all entail the use of mouse strains that synthesize Cre (these strains are designated Cre-deleters), either ubiquitously or in specific tissues or cell types (strain B or C, in the case of *Polb*, reported above), geneticists have agreed to establish a specific database listing all the Cre strains available (The Cre-X-Mice database at http://nagy.mshri.on.ca/cre_new/search/Search.php and The Jackson Laboratory Cre Resources at http://www.creportal.org/). These strains are, in general, freely available on the basis of a material

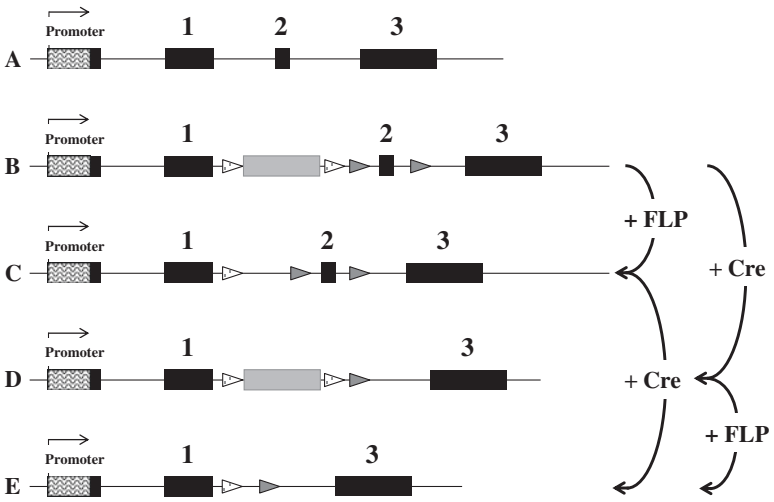


Fig. 8.12 *Inducing targeted deletions with the Cre-loxP and FLP-Frt systems*. The Cre and FLP recombinases can be used successively in the same experiment. In the case presented here, when FLP is used first, the selection cassette (*shaded box*) is deleted ($B \rightarrow C$). Alternatively, if Cre is used first, exon 2 is deleted ($B \rightarrow D$). Finally, when Cre and Frt are used successively, the selection cassette and exon 2 are both deleted ($B \rightarrow E$)

transfer agreement (MTA). This attitude, which is more and more common in the community of mouse geneticists, has saved and still saves a lot of research money. It has been made simpler every day with the use of the internet.

8.3.2.7 Gene Trapping and Targeted Trapping in ES Cells

In an earlier section of this chapter (Sect. 8.3.2.1), we reported experiments in which retroviruses were successfully used for producing insertional mutations in ES cells. These experiments revealed that, unlike the cloned DNA molecules injected into the pronucleus, retroviruses integrate into the genome of ES cells without generating extensive chromosomal rearrangements. In these conditions, when mutations were induced, the proviral copy could be used as a tag for cloning the flanking sequences and finally for identifying the mutated genes. Unfortunately, other than these advantages, the retroviruses have two major drawbacks: first, they insert randomly in the genome and infrequently in exons; second, using a proviral insertion for "harpooning" the flanking sequences is sometimes misleading, especially when there are many proviral insertions in the same ES cells.

In order to improve the efficiency of recovering mutations that are likely to have a phenotypic expression, an original strategy known as *gene trapping* was developed in several laboratories (Gossler et al. 1989; Friedrich and Soriano 1991; Skarnes et al. 1992; Evans et al. 1997; Cecconi and Meyer 2000; Stanford

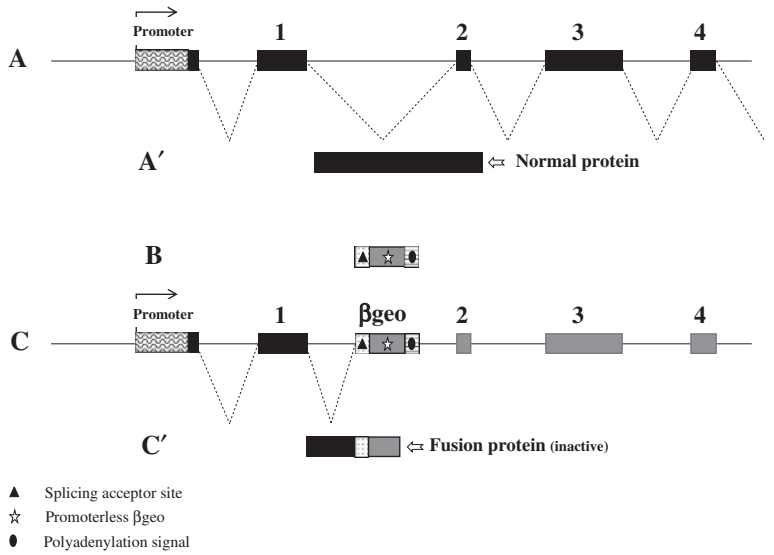


Fig. 8.13 *Gene trapping*. When a promoterless synthetic reporter gene, such as βgeo, sandwiched between a splice acceptor site and a polyadenylation signal (*B*) inserts, by chance, into one of the introns of an expressed gene ($A \rightarrow C$), the reporter gene is transcribed as if it were an exon of the gene. This generates a fusion mRNA, which is (sometimes) translated into a non-functional fusion protein C' (the trapped gene is inactivated). (This figure is redrawn from Skarnes et al. 1992)

et al. 2001; Hansen et al. 2003; Stryke et al. 2003). The principle of this strategy consisted of transfecting ES cells with a promoterless reporter gene and/or a selectable genetic marker flanked, upstream, by a 3′ splice acceptor (SA) site, and downstream by a polyadenylation signal (pA) (Fig. 8.13).

In early experiments, a popular promoterless gene was engineered by fusion of a β-galactosidase moiety (acting as a reporter) with a neomycin-resistant moiety (acting as a selectable marker) and was designated βgeo (contraction of *β-gal* with *neo*). When such a cassette was inserted in an intron, the gene was said to be "trapped." Nowadays, a variety of promoterless artificial genes have been designed with different reporter sequences, making the method more efficient and more versatile.

Transcription of the trapped genes, controlled by the endogenous promoter, resulted in a fusion (or hybrid) RNA molecule, which in turn, was translated into a non-functional protein with some sequence of the endogenous trapped gene beside some others from the sequence of the reporter/selectable marker.[19] Since the encoded fusion protein was non-functional, the trapped genes were equivalent to

[19] Trapping cassettes have also been designed with a marker gene or a selectable gene coupled to a suitable promoter but lacking a downstream polyadenylation signal. In this case, the transcript was also a hybrid molecule, utilizing the 3′ sequences of the host gene to acquire a poly (A) tail.

knockout (or loss-of-function) alleles and the sequence of the cassette could then be used as a tag for gene identification.

Although the strategy of gene-trapping works exclusively with those genes that are transcribed in ES cells, it is nevertheless a high-throughput approach for the identification of genes. It has been (and still is) widely used. Several laboratories, working in an *International Gene-Trap Consortium* (IGTC), have undertaken the establishment of large libraries of ES cells harboring gene-trap insertions. From recent estimates, over 126,500 ES cell lines, each with a trapped gene, are offered to the community on a non-collaborative basis.[20] This represents ~13,300 trapped genes (i.e., around 50 % of all the known genes in the mouse).

In the laboratories performing this type of experiment, the trapped genes are systematically identified unambiguously by using a PCR-based strategy such as 5'RACE (*rapid amplification of cDNA ends*), to generate a sequence tag unique for each insertion. By the way, this is greatly facilitated by the availability of the mouse genome sequence. Researchers who are interested can search and browse the IGTC database (www.genetrap.org) looking for the ES cell lines they are interested in, using accession numbers or IDs, keywords, sequence data, tissue expression profiles, or biological pathways.

As we already mentioned, newer gene-trap vectors have been developed, offering a variety of possibilities for post-insertional modification and the generation of a wide spectrum of alleles.

The trapped-gene libraries that exist nowadays have become an indispensable source of ready-made mutations in mice. For those readers who would like to know more about these libraries, the way they were established and their potential interest we recommend three general publications co-authored by scientists who were deeply involved in their development (Guan et al. 2010; Skarnes et al. 2011; Bradley et al. 2012). The Web site of the International Knockout mouse consortium http://www.knockoutmouse.org/ is also an important source of information, which is user-friendly and explains all the technical steps in the gene-trapping strategy.

As we explained above, gene trapping depends on the random insertion of a reporter cassette in an intron, but the cassette in question can also be inserted in a predefined position by homologous recombination. This strategy is known by the generic name of *targeted trapping* (Friedel et al. 2005). In this case, the vector (basically the same as the one used for gene trapping) is flanked by genomic sequences of the host, completely excluding the promoter. Targeted trapping in mouse ES cells is a simple though powerful tool for analysis of mammalian gene function. Provided the promoterless construct is carefully designed, most random insertions are eliminated by drug selection and the targeting frequencies can reach 50 % or even more (Fig. 8.14).

[20] With, however, some handling fees.

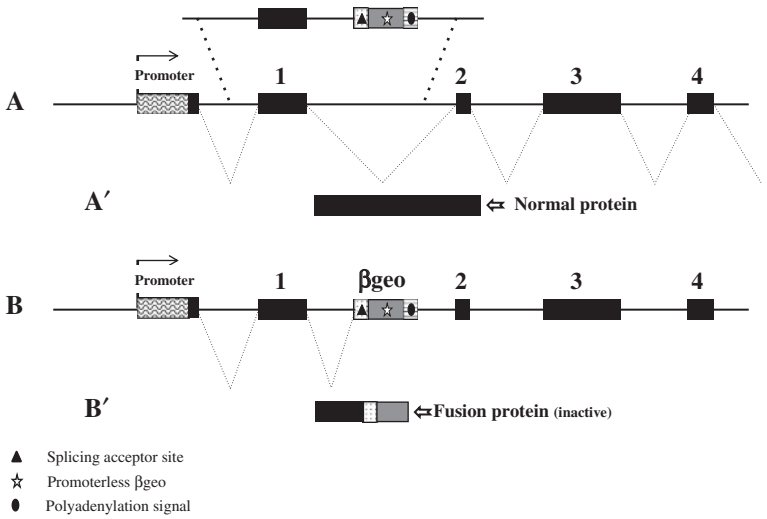


Fig. 8.14 *Targeted trapping*. In this case, insertion of the promoterless reporter gene βgeo is not random, as in the case of gene trapping, but instead results from homologous recombination with a selected region of the targeted gene ($A \rightarrow B$). As in the case of gene trapping, the promoterless gene in the cassette is activated and possibly translated into a fusion protein (B'). In this experiment, it is important that the targeted region does not contain the promoter of the gene. After characterization, the targeted or trapped ES cell clones can be deep-frozen and stored for further use. (This figure is redrawn from Skarnes et al. 1992)

8.3.3 Induction of Mutations in ES Cells with Chemical Mutagens

In Chap. 7, we explained that the induction of mutations in the mouse germ line with radiation or chemical mutagens was an efficient method for the annotation of mammalian genes because it produced all kinds of mutations (nonsense, missense, etc.) and all kinds of alleles (recessive and dominant etc.)—unlike most techniques of ES cell engineering, which produce mostly knockouts (i.e., null alleles). However, a major drawback of chemical mutagenesis is the cost of breeding and/or the time necessary to identify and characterize the new mutations. In addition, all these induced mutations are scattered throughout the whole genome, they are a mixture of different kinds, and they do not necessarily match the interest of the scientist. The genotype-based screens, which consisted of the identification, after analysis performed at the DNA level, of mice heterozygous for a mutation induced by ENU in a specific gene (as described in Chap. 7—Sect. 7.5.4), were considered more advantageous, especially when a deep-frozen sperm bank was available. Unfortunately, here again, this may still be insufficient if a series of alleles at a given locus is desired.

A genotype-based screen for ENU-induced mutations has been adapted with success for the identification of mutations induced in ES cells, in specific genes of interest (Chen et al. 2000; Munroe et al. 2000). In a series of experiments focused on two loci of importance for mouse early development, *Smad2* and *Smad4*, Vivian and colleagues (Vivian et al. 2002) mutagenized 2,060 ES cell clones by incubating the cells for 2 h in a culture medium with 0.2 mg/ml of ENU. They found a total of 29 mutations, out of which 20 were non-silent (yielding a non-silent mutation rate of 1 per 673 kb of screened DNA). This indicates that chemical mutagenesis in mouse ES cells, associated with high-throughput mutation detection, is another interesting method for the identification of mutations in non-selectable genes.

Other experiments on chemical mutagenesis of mouse ES cells have also been suggested as an alternative approach to the chemical mutagenesis of spermatogonia (Becker et al. 2006; Munroe and Schimenti 2009). This strategy has (at least) two major advantages: first, it enables the use of a variety of chemicals with different mutational spectra (different from ENU); second, it allows (at least in theory) the induction of a higher number of mutations in the mouse genome as a consequence of the possibility of performing several successive rounds of mutagenesis in vitro. In addition to these advantages in terms of efficiency, the chemical mutagenesis of ES cells has the same advantages as the gene-driven strategy described in Chap. 7 that it requires only two generations of breeding to reveal the phenotype of the induced mutations (breeding G1s, then intercrossing the G1). In addition, just like for the sperm cells in the case of gene-driven strategy, samples of successfully treated ES cells can be stored deep-frozen as long as necessary for the further detection of induced mutations. This method has not been used very much, probably because the techniques of genetic engineering were developed concurrently, but their advantages, as outlined above, are unique and should be kept in mind.

8.4 Inducible Transgenesis: The *Tet-off* and *Tet-on* Expression Systems

The Cre-*lox*P and the FLP-*Frt* strategies allow the induction of conditional gene knockout. With these strategies, researchers can inactivate virtually any gene, in any specific tissue or cell lineage, and when desired. However, once the Cre-recombinase has excised a floxed DNA segment, the situation is irreversible: the gene is permanently inactivated (or activated) in all daughter cells. Obviously, this may represent a drawback in experiments where only a transient inactivation (or activation) would be desired. It also may be desirable, in some experiments with transgenic mice, to have a transgene expressed only during a certain period but switched off the rest of the time. Unfortunately, this is not possible with the techniques described above.

The *Tet-off* and *Tet-on* inducible expression systems overcome these problems, placing the transcription of a given transgene under the control of the researcher.

In this system, the expression of a transgene is dependent on a tetracycline-controlled transactivator protein and can be regulated, both reversibly and quantitatively, by exposing the transgenic mice to the antibiotic tetracycline (Tc) or to one of its derivatives such as *doxycycline* (Dox). The technology was developed by Bujard and colleagues at the University of Heidelberg (Gossen and Bujard 1992; Baron and Bujard 2000).

The *Tet-off* system requires two critical ingredients. The first is the tetracycline-controlled transactivator protein (in short tTA). tTA is an artificial protein created

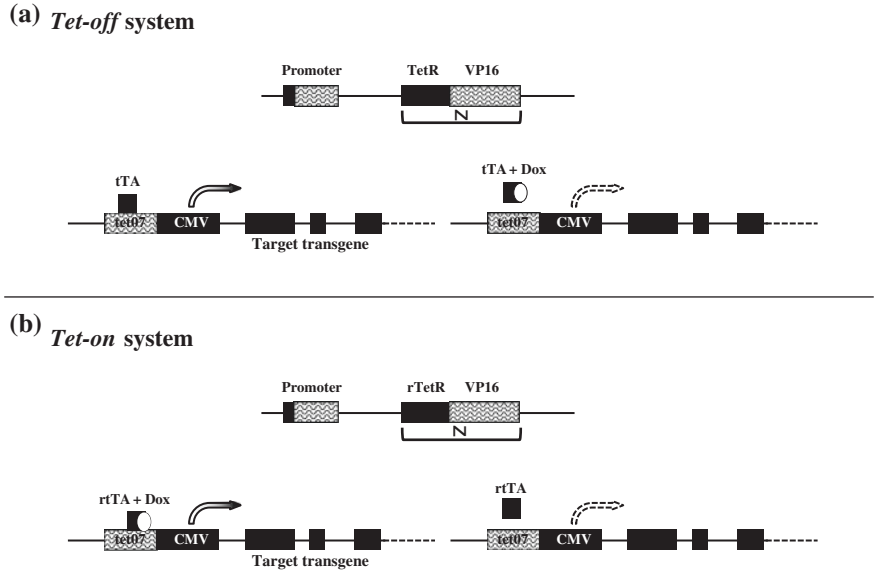


Fig. 8.15 *The "Tet-off" and "Tet-on" Expression Systems*. The *Tet-off* and *Tet-on* inducible expression systems enable transgene expression to be dependent on a tetracycline-controlled transactivator protein (tTA). Under these conditions, transgenic expression can be regulated. **a** The *Tet-off* system requires two ingredients. The first is the tTA, which is a fusion protein created with the TetR (tetracycline repressor), found in *Escherichia coli* transposon Tn10 and encoding resistance to the antibiotic tetracyclin, and a strong trans-activating domain of an herpes simplex virus protein called VP16. The second ingredient is the tetracycline-responsive promoter element (*TRE*) that is composed of a concatemer of seven *tet* operators (tetO7) fused to the minimal promoter sequences of the human cytomegalovirus immediate early gene 1 (*hCMVIE1*) promoter/enhancer. In the absence of tetracyclin (*Tc*) or doxycyclin (*Dox*), tTA binds to TRE and activates expression of the targeted gene. This induction returns to basal levels or is suppressed upon administration of Tc or Dox. The *Tet-on* system works in exactly the opposite manner. This system is based on a reverse tetracycline-controlled trans-activator (*rtTA*), which is also a fusion protein composed of the TetR and the VP16 transactivation domain. However, a four amino acid change in the TetR DNA-binding moiety alters rtTA's activity binding characteristics in such a way that it can recognize the tetO sequences in the TRE of the target transgene only in the presence of the Dox effector (delivered in the water or the food). Thus, in the *Tet-on* system, transcription of the TRE-regulated target is stimulated by rtTA only in the presence of Dox. **b** As explained in the text, both systems require the generation of double transgenic (or bigenic) mice carrying, in the same genome, the TRE-regulated target transgene and the tetracycline-controlled transactivator (*tTA* or *rtTA*)

by fusion of the TetR (tetracycline repressor), found in *Escherichia coli* transposon Tn10 and encoding resistance to Tc, with a strong transactivating domain of the herpes simplex virus protein called VP16. The second critical ingredient required for the *Tet-off* system to operate is the Tc-responsive promoter element (TRE). This promoter is composed of a concatamer of seven *tet* operators (tetO7) fused to the minimal promoter sequences of the human cytomegalovirus immediate early gene 1 (*hCMVIE1*) promoter/enhancer. In the absence of Tc or Dox, tTA binds to TRE and activates expression of the target gene. This induction returns to basal levels or is suppressed upon administration of Tc or Dox (Fig. 8.15).

The *Tet-on* system works in exactly the opposite manner. It is based on a reverse tetracycline-controlled transactivator (rtTA), which is also a fusion protein composed of TetR and the VP16 transactivation domain; however, a four-amino-acid change in the TetR DNA-binding moiety alters rtTA's activity binding characteristics such that it can recognize the tetO sequences in the TRE of the target transgene only in the presence of the Dox effector. Thus, in the *Tet-on* system, transcription of the TRE-regulated target is stimulated by rtTA only in the presence of Dox (i.e., when the drug is delivered either in the drinking water or with the food). A good example of the value of this system for cancer research is a model where an activated *Kras* oncogene is inducibly expressed in an epithelial compartment using keratin 5 (K5)-rtTA: tet-Kras bigenic mice (Vitale-Cross et al. 2004).

These *Tet-off* and *Tet-on* systems can be used, for example, to design dominant gain-of-function experiments in which temporal control of transgene expression is required (Gossen and Bujard 1992; Furth et al. 1994; Kistner et al. 1996; Schonig and Bujard 2003). The *Tet-off* expression system is more popular than the *Tet-on* system because it does not require the constant administration of a drug whose effects might be deleterious in the longterm.

8.5 Other Techniques for the Production of Transgenic Mice

Considering the efficiency of ENU mutagenesis and the potentialities of genetic engineering applied to ES cells, it is clear that mouse geneticists have at their disposition an unmatched arsenal of strategies allowing them to generate virtually any type of alteration in the genome of their favorite species. This is unfortunately not the case with other species of mammals, especially the rat, which is yet another important source of model for human diseases[21]. However, techniques have been developed to generate genomic alterations in these species and some have proved very promising. Most of these techniques have been efficiently and successfully transposed to the mouse species. We will describe the most important.

[21] Some domestic species (the rat in particular), present phenotypes that have not yet been documented in the mouse; this is why it would be important that the genetic arsenal that has been developed for the mouse be replicated in these other species.

8.5.1 Transgenesis by Retroviral Infection of Early Embryos

The integration of exogenous DNA into the germ line through experimental infection of mouse embryos with retroviruses was successfully achieved a long time ago (Jaenisch 1976). Newborns and preimplantation embryos (4–8 cell stage) were infected with the Moloney murine leukemia virus (M-MuLV), and it was observed that infection of preimplantation embryos, in contrast to infection of newborns, could lead to stable integration of proviral copies into the germline. These initial experiments have yielded several mouse strains with stable germ line integrations of retroviral DNA at distinct chromosomal loci (for example, the *Mov* loci; Jaenisch 1976). One of these integrations was in the gene encoding procollagen, type I, alpha 1 ($Col1a1^{Mov13}$) (Stacey et al. 1988).

Experimental infections of preimplantation embryos have the advantage that the viral integrations are in general stable and do not generate the sort of chromosomal rearrangements that often occur with the classical pronuclear techniques. Since these integrations occur almost at random, they sometimes hit a gene (as in the case of *Col1a1*) and produce a visible mutant phenotype. Here again, the DNA of the retrovirus can be used as a "hook" to clone the DNA sequences flanking the insertion site, and this helps in the characterization of the mutant allele.

Viral infection can also be used to introduce foreign DNA into embryos or eukaryotic cells in culture, and the advantages of using mouse retroviruses as shuttles for transgenesis have been explained in detail in a review by Nicolas and Rubenstein (1988). Two of these advantages are noteworthy in the context of this chapter:

- All the sequences of the viral genome required for its replication, transcription, and integration are grouped in or adjacent to the long terminal repeat (LTR).
- All the necessary proteins for infection, reverse transcription, and integration of the viral genome can be removed from the "shuttle" virus and provided in trans by a "helper" virus, leaving space for foreign DNA inserts of up to 8–10 kb.

For transgenesis in rodents (mostly in rat), the lentiviruses derived from human HIV have been the most widely employed (Wiznerowicz and Trono 2005). The reason for this choice is that lentiviruses, unlike most other retroviruses, have the capacity to infect nondividing cells. Shuttle viruses are produced by transfection of the construct into packaging cell lines, which are engineered to provide the essential viral proteins for assembly of infectious particles. The viruses are harvested from the cell culture medium and used for microinjection into the perivitelline space of single-cell embryos (Koentgen et al. 2010). Infected embryos reverse-transcribed the lentiviral RNA into DNA (provirus) that inserts back into the genome. However, because they are defective, the viruses are capable of completing only a single infectious cycle but cannot replicate further.

Lentiviral integrations, in addition to being relatively stable and because they are less invasive than pronuclear injections, sometimes yield survival rates approaching 90 %. Another advantage is that lentiviruses integrate as single copies and are expressed more reliably than the transgenes obtained by pronuclear injections; in particular, they are less prone to epigenetic silencing (Koentgen et al. 2010). The major weakness of this technique is the limit of 8–10 kb for the transgene size.

8.5.2 *In Vivo Genome Editing: The Production of Targeted Alterations Using Engineered Nucleases*

Over the last 10 years, a totally new kind of technique has been developed for the production of gene- (or locus-) targeted mutations that make use of engineered hybrid molecules which associate sequence-specific DNA-binding domains with a non-specific DNA cleavage domain. These techniques have demonstrated significant advantages for the production of a variety of mutations at targeted sites in several species commonly used by geneticists, including *Arabidopsis thaliana*, *Caenorhabditis elegans*, the sea urchin *Echinus melo*, *Drosophila melanogater*, and *Danio rerio*, to cite just a few. Recently, the techniques in question have been successfully adapted to the production of targeted mutations (knockout and knock-in) in mammals, mainly in the rat (Geurts et al. 2009), the mouse (Carbery et al. 2010), and other domestic species (reviewed in Rémy et al. 2010; Gaj et al. 2013; Kim and Kim 2014; and Mashimo 2014). We will describe some of these techniques and discuss their possible applications for genome editing.

8.5.2.1 Zinc-Finger Nucleases and Transcription Activator-like Effector Nucleases

The molecules used in initial experiments associated zinc-finger DNA-binding motifs with the restriction endonuclease *Fok*I, and for this reason, they were called *zinc-finger nucleases* (ZFNs). For the production of mutations, two complementary ZFNs must be designed, each of them recognizing a specific DNA sequence spanning 9–18 bp on either side of a 5–6-bp sequence defining the targeted region. When injected into a cell or a pronucleus, the ZFNs assemble tightly on both sides of the targeted site, one on each strand, and *Fok*I performs double-strand breaks (DSBs).[22] Once cleaved by the endonuclease, the cellular mechanisms controlling DNA integrity are immediately triggered to repair the damage. These mechanisms are of two types. The first is known as the homology-dependent repair (HDR) mechanism, which requires a homologous (template) sequence to guide the repair: it is precise and accurate and re-establishes *ad integrum* the original sequence of the cleaved DNA strand. The second mechanism, known as non-homologous end joining (NHEJ), is more common but is more rapidly activated. NHE is much less precise and only approximately restores the damaged strands, leaving behind deletions of nucleotides and accordingly frame shift mutations that are in most instances loss-of-function mutations (Fig. 8.16).

[22] A specific ZFN binds with 3 bp at the DNA level. Since there is a great variety of such motifs, a judicious selection of 3–6 of them allows the targeting of a 9–18-bp DNA domain, which is highly specific. Libraries of ready-made ZFNs are also available which allow the targeting of virtually any sequence in the mouse genome.

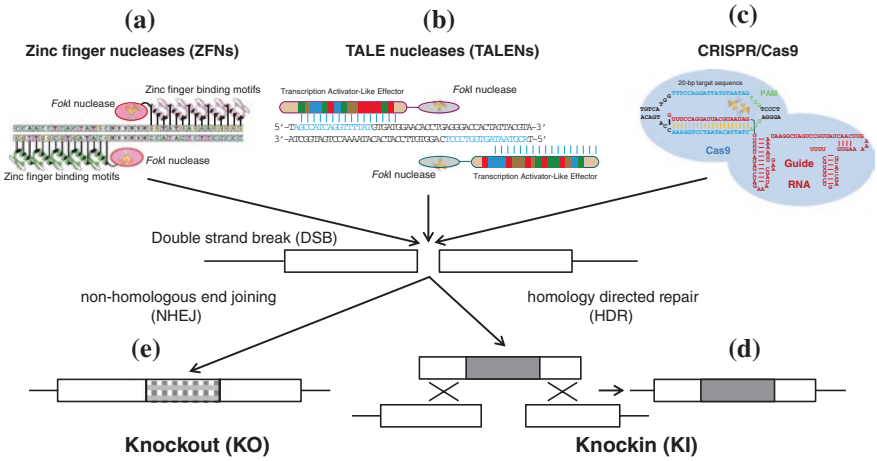


Fig. 8.16 *The production of targeted genome alterations using site-specific engineered nucleases or the CRISPR/Cas9-based RNA-guided DNA endonuclease*. The figure schematically represents three strategies used for genome editing. Zinc fingers (**a**) and TALEN modules (**b**) both bind to adjacent DNA sequences in opposite directions, leaving a small gap in between for the *Fok*I endonuclease to perform a double-stranded break (DSB). **c** With the CRISPR strategy, Cas9 unwinds the DNA duplex and performs a DSB after recognition of a specific (~20 bp) target by the gRNA, provided that the correct protospacer adjacent motif (PAM) is present. Whatever their origins, DSBs are ligated through non-homologous end joining (NHEJ) (**d**) or repaired through homology-directed repair (*HDR*) (**e**). For HDR to occur, a DNA molecule or a single-stranded synthetic DNA must be added as a template. If the sequence of the template differs from the endogenous sequence by the addition or substitution of some nucleotides, this results in a knock-in. These methods for producing mutations at specifically targeted sites are very efficient. The CRISPR/Cas9-based RNA-guided strategy permits the production of several independent point mutations in the same genome (Courtesy T. Mashimo)

The technique is simple in its practical aspects. Messenger RNAs transcribed in vitro from engineered ZFN plasmids are injected into the (male) pronuclei of mouse zygotes, exactly as in the case of pronuclear (*in ovo*) transgenesis, then the embryos are transferred into the oviduct of pseudo-pregnant females. With this technique a homozygous knockout mutation can be obtained in 4–5 months, which is much faster than with the traditional knockout strategies using ES cells.[23] Another important advantage of the technique is that it is applicable to all strains of mice, allowing for the production of a series of mutations at the same locus in different inbred backgrounds (co-isogenic strains) (Carbery et al. 2010). Finally, the technique can produce a variety of mutations, mainly deletions ranging from 1 bp to more than 1 kb, and more rarely, insertions of a few bp, but also sequence-specific mutations—which are all potential tools for the analysis of the targeted

[23] This comment concerning the time necessary to produce a knockout mutation in the mouse genome by using the ZFN strategy, although reduced, must nevertheless be compared with the time necessary to purchase, when available, an ES cell line harboring the same ready-made knockout, when the latter is available in a repository such as KOMP (https://www.komp.org/).

gene's function. Knock-in mice and rats carrying sequence-specific modifications have already been produced using ZFN technology (Cui et al. 2011).

The expression of artificial nucleases in embryonic cells at early stages of development does not seem to be toxic or to have any breakage activity outside of the targeted DNA sequence (little or no off-target events). The only drawback of the technique, which is not a major one, is that alterations may still occur at the targeted site several days after the injection, making some founder animals behave like mosaics. Another potential drawback is that, for technical reasons, the technique is probably not applicable to gene families, since the design of sequence-specific domains of the ZFNs would be difficult or even impossible in this case.

This basic technique of genome modification making use of ZFNs has undergone several improvements and developments. The first one is based on the observation, already mentioned above, that when DSBs are induced in cells by any means (for example, as a consequence of irradiation or of nuclease activity—and regardless of the nuclease) the homology-dependent repair mechanism (HDR) is activated. These mechanisms increase the potentialities of insertion of exogenous DNA that have sequence homologies at their ends with the sequence flanking the DSB. For example, adding the cloned RNA of the *lac*Z reporter gene to the mRNAs injected into the pronucleus allowed the production of knock-in transgenic mice with *lacZ* integrated between the boundaries of the DSB.

Another improvement has been the replacement of the DNA-binding components of the ZFNs by molecules from the plant bacterium *Xanthomonas* with similar DNA-binding properties. These molecules are a family of transcription activator-like effectors (TALEs) and the DNA-binding hybrid molecules are known as TALENs. TALEs have binding capacities greater than ZFNs and can match with virtually any sequence, further increasing the efficiency of the technique (Tesson et al. 2011). Over recent years, several groups have used TALENs to modify endogenous genes in a wide variety of species including insects, amphibians, fish (zebrafish), and mammals (rat, mouse, pig, and cow) (Joung and Sander 2013; Sung et al. 2013). The advantages of TALENs over ZFNs include their ease of design and assembly, their specificity, and their lower cost. Injection of the exonuclease *Exo1* in substitution for endonuclease *FokI* in the TALEN technique has been another improvement in the production of knockouts in rats (Mashimo et al. 2013).

8.5.2.2 The CRISPR/Cas9 System

The strategies that we described in the section above consisted of the production of double-strand breaks (DSBs) by the protein-guided DNA cleavage activity of engineered ZFNs or TALENs. Recently, another technique has been developed that depends on small RNAs for the production of sequence-specific cleavages (RNA-guided DNA cleavage). This strategy was developed after the identification and characterization of a defense mechanism, known as the CRISPR/Cas system,

which operates in bacteria and archaea and allows these organisms to fight infections by viruses, plasmids, or phages (Pennisi 2013).[24]

A CRISPR locus consists of a series of short direct repeats (average size 32 bp) of identical sequences, interspersed with intervening regions called *spacers*, which consist of small but variable sequences. Analysis of the sequence of these spacers indicates great similarities with the sequences of some phages and plasmids, providing a possible interpretation for the mechanism of recognition of the genome of the invaders by the CRISPR.

The CRISPR loci are transcribed into short CRISPR RNAs (crRNA). These crRNAs anneal to transactivating crRNAs (tracrRNAs) and direct sequence-specific cleavage of DNA by Cas proteins. Target recognition by the CRISPR-associated nuclease (Cas9) protein requires a *seed* sequence within the crRNA and a conserved dinucleotide-containing *protospacer adjacent motif* (PAM) sequence upstream of the crRNA-binding region (Fig. 8.16).

Engineered modifications of the CRISPR, as well as the Cas9 part, have led to an efficient way of producing DSBs at will. The CRISPR component is usually referred to as a guide RNA (gRNA). Cas9 utilizes gRNA that binds to specific DNA sequences to produce the DSBs.

The Cas9 protein consists of three more or less independent domains: one DNA-binding domain and two catalytic domains that independently cut one DNA strand. The two domains with nuclease activity can be inactivated separately by simple point mutations, and these modified versions of Cas9, with one cutting domain disabled, introduce single-strand breaks or DNA *nicks*. Even though DNA nicking is less efficient for genome editing, it dramatically reduces the chance of so-called off-target effects, since unwanted nicks are faithfully reconstructed by homology-directed repair (HDR). DSBs can be achieved at the targeted site by a pair of DNA-binding gRNAs, with sites close to each other but on opposite strands.

The RNA-guided endonucleases can be engineered to cleave virtually any DNA sequence by appropriately designing the crRNA; for example, to generate knock-in animals carrying conditional or reporter alleles (Yang et al. 2013). This technique exhibits several advantages over the methods using ZFNs or TALENs. One can, for example, generate mice carrying mutations in multiple genes across the genome in a single step by simultaneously injecting various gRNAs (Horii et al. 2014). This technique is known as multiplex gene editing and has been applied successfully not only to cells cultured in vitro but also to mouse and rat embryos (Wang et al. 2013; Wei et al. 2013). It saves a lot of breeding time when an experimental project requires the presence of several mutations in the same genome.

The genomic alterations that can be produced by using the CRISPR/Cas9 technology are not limited to the production of indels but can also consist of knock-ins. If we consider that the strategy is relatively easy to apply and somewhat faster than the other strategies using engineered nucleases, we see that CRISPR/Cas9 may well-revolutionize genomic engineering in the near future (Mashimo 2014; Zhang et al. 2014).

[24] CRISPR is an acronym for *clusters of regularly interspaced short palindromic repeats.*

8.6 Conclusion

Contemplating all the many possibilities for creating transgenic mice, one can see that geneticists now have all the tools in hand to answer virtually any questions that may arise in their analysis of gene functions. They also have at their disposition a very large collection of ready-made mutations of all kinds, waiting to be used, for example, as models of human diseases.[25] All these tools and models will be important for performing genome annotation.

References

Abiola O, Angel JM, Avner P, Bachmanov AA, Belknap JK, Bennett B, Blankenhorn EP, Blizard DA, Bolivar V, Brockmann GA, Buck KJ, Bureau JF, Casley WL, Chesler EJ, Cheverud JM, Churchill GA, Cook M, Crabbe JC, Crusio WE, Darvasi A, de Haan G, Dermant P, Doerge RW, Elliot RW, Farber CR, Flaherty L, Flint J, Gershenfeld H, Gibson JP, Gu J, Gu W, Himmelbauer H, Hitzemann R, Hsu HC, Hunter K, Iraqi FF, Jansen RC, Johnson TE, Jones BC, Kempermann G, Lammert F, Lu L, Manly KF, Matthews DB, Medrano JF, Mehrabian M, Mittlemann G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Mountz JD, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Paigen B, Palmer AA, Peirce JL, Pomp D, Rosemann M, Rosen GD, Schalkwyk LC, Seltzer Z, Settle S, Shimomura K, Shou S, Sikela JM, Siracusa LD, Spearow JL, Teuscher C, Threadgill DW, Toth LA, Toye AA, Vadasz C, Van Zant G, Wakeland E, Williams RW, Zhang HG, Zou F; Complex Trait Consortium. (2003) The nature and identification of quantitative trait loci: a community's view. Nature Review Genetics 4:911–916

Adams JM, Harris AW, Pinkert CA, Corcoran LM, Alexander WS, Cory S, Palmiter RD, Brinster RL (1985) The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. Nature 318:533–538

Babinet C, Cohen-Tannoudji M (2001) Genome engineering via homologous recombination in mouse embryonic stem (ES) cells: an amazingly versatile tool for the study of mammalian biology. Anais da Academia Brasileira de Ciencias 73:365–383

Ballester M, Castelló Anna, Ibáñez E, Sánchez A, Folch JM (2004) Real-time quantitative PCR-based system for determining transgene copy number in transgenic animals. BioTechniques 37:610–613

Baron U, Bujard H (2000) Tet repressor-based system for regulated gene expression in eukaryotic cells: principles and advances. Methods Enzymol 327:401–421

Becker S, de Angelis MH, Beckers J (2006) Use of chemical mutagenesis in mouse embryonic stem cells. Methods Mol Biol 329:397–407

Bradley A, Anastassiadis K, Ayadi A, Battey JF, Bell C, Birling MC, Bottomley J, Brown SD, Bürger A, Bult CJ, Bushell W, Collins FS, Desaintes C, Doe B, Economides A, Eppig JT, Finnell RH, Fletcher C, Fray M, Frendewey D, Friedel RH, Grosveld FG, Hansen J, Hérault Y, Hicks G, Hörlein A, Houghton R, Hrabé de Angelis M, Huylebroeck D, Iyer V, de Jong PJ, Kadin JA, Kaloff C, Kennedy K, Koutsourakis M, Lloyd KC, Marschall S, Mason J, McKerlie C, McLeod MP, von Melchner H, Moore M, Mujica AO, Nagy A, Nefedov M, Nutter LM, Pavlovic G, Peterson JL, Pollock J, Ramirez-Solis R, Rancourt DE, Raspa M,

[25] To paraphrase the title of an interesting review on the subject one could say that, nowadays, geneticists have at their disposition "*a mouse for all reasons*" (International Mouse Knockout Consortium 2007).

Remacle JE, Ringwald M, Rosen B, Rosenthal N, Rossant J, Ruiz Noppinger P, Ryder E, Schick JZ, Schnütgen F, Schofield P, Seisenberger C, Selloum M, Simpson EM, Skarnes WC, Smedley D, Stanford WL, Stewart AF, Stone K, Swan K, Tadepally H, Teboul L, Tocchini-Valentini GP, Valenzuela D, West AP, Yamamura K, Yoshinaga Y, Wurst W (2012) The mammalian gene function resource: the International Knockout Mouse Consortium. Mamm Genome 23:580–586

Breitman ML, Clapoff S, Rossant J, Tsui LC, Glode LM, Maxwell IH, Bernstein A (1987) Genetic ablation: targeted expression of a toxin gene causes microphthalmia in transgenic mice. Science 238:1563–1565

Breitman ML, Bryce DM, Giddens E, Clapoff S, Goring D, Tsui LC, Klintworth GK, Bernstein A (1989) Analysis of lens cell fate and eye morphogenesis in transgenic mice ablated for cells of the lens lineage. Development 106:457–463

Brinster RL, Chen HY, Trumbauer M, Senear AW, Warren R, Palmiter RD (1981) Somatic expression of herpes thymidine kinase in mice following injection of a fusion gene into eggs. Cell 27:223–231

Brinster RL, Allen JM, Behringer RR, Gelinas RE, Palmiter RD (1988) Introns increase transcriptional efficiency in transgenic mice. Proc Natl Acad Sci USA 85:836–840

Carbery ID, Ji D, Harrington A, Brown V, Weinstein EJ, Liaw L, Cui X (2010) Targeted genome modification in mice using zinc-finger nucleases. Genetics 186:451–459

Capecchi MR (1989) Altering the genome by homologous recombination. Science 244:1288–1292

Cecconi F, Meyer BI (2000) Gene trap: a way to identify novel genes and unravel their biological function. FEBS Lett 480:63–71

Chen Y, Yee D, Dains K, Chatterjee A, Cavalcoli J, Schneider E, Om J, Woychik RP, Magnuson T (2000) Genotype-based screen for ENU-induced mutations in mouse embryonic stem cells. Nat Genet 24:314–317

Costantini F, Lacy E (1981) Introduction of a rabbit beta-globin gene into the mouse germline. Nature 294:92–94

Cui X, Ji D, Fisher DA, Wu Y, Briner DM, Weinstein EJ (2011) Targeted integration in rat and mouse embryos with zinc-finger nucleases. Nat Biotechnol 29:64–67

DeChiara TM (2001) Gene targeting in ES cells. Methods Mol Biol 158:19–45

Dorner M, Horwitz JA, Robbins JB, Barry WT, Feng Q, Mu K, Jones CT, Schoggins JW, Catanese MT, Burton DR, Law M, Rice CM, Ploss A (2011) A genetically humanized mouse model for hepatitis C virus infection. Nature 474:208–211

Duboule D (1998) Vertebrate Hox gene regulation: clustering and/or colinearity? Curr Opin Gene Develop 8:514–518

Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. Nature 292:154–156

Evans MJ, Carlton MB, Russ AP (1997) Gene trapping and functional genomics. Trends Genet 13:370–374

Feil S, Valtcheva N, Feil N (2009) Inducible Cre mice methods. Mol Biol 530:343–363

Filippov MA, Hormuzdi SG, Fuchs EC, Monyer H, Robertson E, Bradley A, Kuehn M, Evans M (2003) A reporter allele for investigating connexin 26 gene expression in the mouse brain. Eur J Neurosci 18:3183–3192

Friedel RH, Plump A, Lu X, Spilker K, Jolicoeur C, Wong K, Venkatesh TR, Yaron A, Hynes M, Chen B, Okada A, McConnell SK, Rayburn H, Tessier-Lavigne M (2005) Gene targeting using a promoterless gene trap vector ("targeted trapping") is an efficient method to mutate a large fraction of genes. Proc Natl Acad Sci U S A 102:13188–13193

Friedrich G, Soriano P (1991) Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. Genes Dev 5:1513–1523

Furth PA, St Onge L, Böger H, Gruss P, Gossen M, Kistner A, Bujard H, Hennighausen L (1994) Temporal control of gene expression in transgenic mice by a tetracycline-responsive promoter. Proc Natl Acad Sci USA 91:9302–9306

Gaj T, Gersbach CA, Barbas CF III (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31:397–405

Geurts AM, Cost GJ, Freyvert Y, Zeitler B, Miller JC, Choi VM, Jenkins SS, Wood A, Cui X, Meng X, Vincent A, Lam S, Michalkiewicz M, Schilling R, Foeckler J, Kalloway S, Weiler H, Ménoret S, Anegon I, Davis GD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Jacob HJ, Buelow R (2009) Knockout rats via embryo microinjection of zinc-finger nucleases. Science 325:433

Gomes-Pereira M, Cooper TA, Gourdon G (2011) Myotonic dystrophy mouse models: towards rational therapy development. Trends Mol Med 17:506–517

Gordon JW, Ruddle FH (1981) Integration and stable germline transmission of genes injected into mouse pronuclei. Science 214:1244–1246

Goring DR, Rossant J. Clapoff S, Breitman ML, Tsui LC (1987) In situ detection of beta-galactosidase in lenses of transgenic mice with a gamma-crystallin/lacZ gene. Science 235:456–458

Gossen M, Bujard H (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci USA 89:5547–5551

Gossler A, Doetschman T, Korn R, Serfling E, Kemler R (1986) Transgenesis by means of blastocystderivedembryonic stem cell lines. Proc Natl Acad Sci USA 83:9065–9069

Gossler A, Joyner AL, Rossant J, Skarnes WC (1989) Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science 244:463–465

Gridley T, Gray DA, Orr-Weaver T, Soriano P, Barton DE, Francke U, Jaenisch R (1990) Molecular analysis of the Mov 34 mutation: transcript disrupted by proviral integration in mice is conserved in *Drosophila*. Development 109:235–242

Gu H, Marth JD, Orban PC, Mossmann H, Rajewsky K (1994) Deletion of a DNA polymerase betagene segment in T cells using cell type-specific gene targeting. Science 265:103–106

Guan C, Ye C, Yang X, Gao J (2010) A review of current large-scale mouse knockout efforts. Genesis 48:73–85

Hammes A, Schedl A (2000) Generation of transgenic mice from plasmids, BACs and YACs. In: Jackson IJ, Abbott CM (eds) Mouse genetics and transgenesis: a practical approach. Oxford University Press, New York, pp 217–245

Hanahan D, Wagner EF, Palmiter RD (2007) The origins of oncomice: a history of the first transgenic mice genetically engineered to develop cancer. Genes Dev 21:2258–2270

Hansen J, Floss T, Van Sloun P, Fuchtbauer EM, Vauti F, Arnold HH, Schnutgen F, Wurst W, von Melchner H, Ruiz P (2003) A large-scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. Proc Natl Acad Sc USA 100:9918–9922

Harbers K, Jahner D, Jaenisch R (1981) Microinjection of cloned retroviral genomes into mouse zygotes: integration and expression in the animal. Nature 293:540–542

Hasty P, Ramirez-Solis R, Krumlauf R, Bradley A (1991) Introduction of a subtle mutation into theHox-2.6 locus in embryonic stem cells. Nature 350:243–246

Hasty P, Abuin A, Bradley A (2000) Gene targeting, principles, and practice in mammalian cells. In: Joyner AL (ed) Gene targeting: a practical approach. Oxford University Press, New York, pp 1–36

He Y, Chen H, Quon MJ, Reitman M (1995) The mouse obese gene. Genomic organization, promoter activity, and activation by CCAAT/enhancer-binding protein alpha. J Biol Chem 270:28887–28891

Heintz N (2001) BAC to the future: the use of BAC transgenic mice for neuroscience research. Nat Rev Neurosci 2:861–870

Herault Y, Duchon A, Velot E, Maréchal D, Brault V (2012) The in vivo Down syndrome genomic library in mouse. Prog Brain Res 197:169–197

Hitz C, Steuber-Buchberger P, Delic S, Wurst W, Kühn R (2009) Generation of shRNA transgenic mice. Methods Mol Biol 530:101–129

Hogan B, Beddington R, Costantini F, Lacy E (1994) Manipulating the mouse embryo: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Hooper ML (1992) Embryonal stem cells. Harwood Academic, Chur 147 pp
Hooper M, Hardy K, Handyside A, Hunter S, Monk M (1987) HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. Nature 326:292–295
Horii T, Arai Y, Yamazaki M, Morita S, Kimura M, Itoh M, Abe Y, Hatada I (2014) Validation of microinjection methods for generating knockout mice by CRISPR/Cas-mediated genome engineering. Sci Rep 4:4513. doi:10.1038/srep04513
Houdebine LM (2003) Animal transgenesis and cloning. Wiley, New York
Huxley C, Passage E, Manson A, Putzu G, Figarella-Branger D, Pellissier JF, Fontes M (1996) Construction of a mouse model of Charcot-Marie-Tooth disease type 1A by pronuclear injection of human YAC DNA. Hum Mol Genet 5:563–569
International Mouse Knockout Consortium, Collins FS, Rossant J, Wurst W (2007) A mouse for all reasons. Cell 128:9–13
Jackson IJ, Abbott CM (2000) Mouse genetics and transgenesis: a practical approach. Oxford University Press, Oxford
Jaenisch R (1976) Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. Proc Natl Acad Sci USA 73:1260–1264
Jaenisch R (1988) Transgenic animals. Science 240:1468–1474
Jaenisch R, Jahner D, Nobis P, Simon I, Lohler J, Harbers K, Grotkopp D (1981) Chromosomal position and activation of retroviral genomes inserted into the germline of mice. Cell 24:519–529
Jakobovits A, Moore AL, Green LL, Vergara GJ, Maynard-Currie CE, Austin HA, Klapholz S (1993) Germline transmission and expression of a human-derived yeast artificial chromosome. Nature 362:255–258
Joung JK, Sander JD (2013) TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol 14:49–55
Katsuki M, Sato M, Kimura M, Yokoyama M, Kobayashi K, Nomura T (1988) Conversion of normal behavior to shiverer by myelin basic protein antisense cDNA in transgenic mice. Science 241:593–595
Kim H, Kim JS (2014) A guide to genome engineering with programmable nucleases. Nat Rev Genet 15:321–334
Kistner A, Gossen M, Zimmermann F, Jerecic J, Ullmer C, Lübbert H, Bujard H (1996) Doxycycline-mediated quantitative and tissue-specific control of gene expression in transgenic mice. Proc Natl Acad Sci USA 93:10933–10938
Koentgen F, Suess G, Naf D (2010) Engineering the mouse genome to model human disease for drug discovery. Methods Mol Biol 602:55–77
Koike S, Taya C, Kurata T, Abe S, Ise I, Yonekawa H, Nomoto A (1991) Transgenic mice susceptible to poliovirus. Proc Natl Acad Sci USA 88:951–955
Kuehn MR, Bradley A, Robertson EJ, Evans MJ (1987) A potential animal model for Lesch–Nyhan syndrome through introduction of HPRT mutations into mice. Nature 326:295–298
Kuhn R, Schwenk F, Aguet M, Rajewsky K (1995) Inducible gene targeting in mice. Science 269:1427–1429
Lakso M, Sauer B, Mosinger B Jr, Lee EJ, Manning RW, Yu SH, Mulder KL, Westphal H (1992) Targeted oncogene activation by site-specific recombination in transgenic mice. Proc Natl Acad Sci USA 89:6232–6236
Lecuit M, Vandormael-Pournin S, Lefort J, Huerre M, Gounon P, Dupuy C, Babinet C, Cossart P (2001) A transgenic model for listeriosis: role of internalin in crossing the intestinal barrier. Science 292:1722–1725
Lee JT, Jaenisch R (1996) A method for high efficiency YAC lipofection into murine embryonic stem cells. Nucleic Acids Res 24:5054–5055
Lichtman JW, Livet J, Sanes JR (2008) A technicolour approach to the connectome. Nat Rev Neurosci 9:417–422
Lira SA, Kinloch RA, Mortillo S, Wassarman PM (1990) An upstream region of the mouse ZP3 gene directs expression of firefly luciferase specifically to growing oocytes in transgenic mice. Proc Natl Acad Sci USA 87:7215–7219

Lithner CU, Hedberg MM, Nordberg A (2011) Transgenic mice as a model for Alzheimer's disease. Curr Alzheimer Res 8:818–831

Lowe LA, Yamada S, Kuehn MR (2001) Genetic dissection of nodal function in patterning the mouse embryo. Development 128:1831–1843

Martin GR, Stevens ME, Bissada N, Nasir J, Kanazawa I, Disteche CM, Rubin EM, Hayden MR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci USA 78:7634–7638

Martin N, Jaubert J, Gounon P, Salido E, Haase G, Szatanik M, Guénet JL (2002) A missense mutation in Tbce causes progressive motor neuronopathy in mice. Nat Genet 32:443–447

Mashimo T (2014) Gene targeting technologies in rats: zinc finger nucleases, transcription activator-like effector nucleases, and clustered regularly interspaced short palindromic repeats. Dev Growth Differ 56:46–52

Mashimo T, Kaneko T, Sakuma T, Kobayashi J, Kunihiro Y, Voigt B, Yamamoto T, Serikawa T (2013) Efficient gene targeting by TAL effector nucleases coinjected with exonucleases in zygotes. Sci Rep 3:1253. doi:10.1038/srep01253

Meisler MH (1992) Insertional mutation of 'classical' and novel genes in transgenic mice. Trends Genet 8:341–344

Messing A, Behringer RR, Slapak JR, Lemke G, Palmiter RD, Brinster RL (1990) Insertional mutation at the *ld* locus (again!) in a line of transgenic mice. Mouse Genome 87:107

Misteli T, Spector D (1997) Applications of the green fluorescent protein in cell biology and biotechnology. Nat Biotechnol 15:961–964

Munroe RJ, Bergstrom RA, Zheng QY, Libby B, Smith R, John SW, Schimenti KJ, Browning VL, Schimenti JC (2000) Mouse mutants from chemically mutagenized embryonic stem cells. Nat Genet 24:318–321

Munroe RJ, Schimenti JC (2009) Mutagenesis of mouse embryonic stem cells with ethylmethanesulfonate. Methods Mol Biol 530:131–138

Nagy A (2000) Cre recombinase: the universal reagent for genome tailoring. Genesis 26:99–109

Nagy A, Gertsenstein M, Vintersten K, Behringer R (2003) Manipulating the mouse embryo, a laboratory manual, 3rd edn. Cold Spring Harbor Press, New York

Nicolas JF, Rubenstein JL (1988) Retroviral vectors. Biotechnology 10:493–513

O'Doherty A, Ruf S, Mulligan C, Hildreth V, Errington ML, Cooke S, Sesay A, Modino S, Vanes L, Hernandez D, Linehan JM, Sharpe PT, Brandner S, Bliss TV, Henderson DJ, Nizetic D, Tybulewicz VL, Fisher EM (2005) An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. Science 309:2033–2037

Overbeek PA, Chepelinsky AB, Khillan JS, Piatigorsky J, Westphal H (1985) Lens-specific expression and developmental regulation of the bacterial chloramphenicol acetyltransferase gene driven by the murine alpha A-crystallin promoter in transgenic mice. Proc Natl Acad Sci USA 82:7815–7819

Overbeek PA, Gorlov IP, Sutherland RW, Houston JB, Harrison WR, Boettger-Tong HL, Bishop CE, Agoulnik AI (2001) A transgenic insertion causing cryptorchidism in mice. Genesis 30:26–35

Palmiter RD, Brinster RL, Hammer RE, Trumbauer ME, Rosenfeld MG, Birnberg NC, Evans RM (1982) Dramatic growth of mice that develop from eggs microinjected with metallothionein–growth hormone fusion genes. Nature 300:611–615

Papaioannou VE, McBurney M, Gardner RL, Evans MJ (1975) The fate of teratocarcinoma cells injected into early mouse embryos. Nature 258:70–73

Passamaneck YJ, Di Gregorio A, Papaioannou VE, Hadjantonakis AK (2006) Live imaging of fluorescent proteins in chordate embryos: from ascidians to mice. Microsc Res Tech 69:160–167

Pennisi E (2013) The CRISPR craze. Science 341:833–836

Pereira R, Khillan JS, Helminen HJ, Hume EL, Prockop DJ (1993) Transgenic mice expressing a partially deleted gene for type I procollagen (COL1A1). A breeding line with a phenotype of spontaneous fractures and decreased bone collagen and mineral. J Clin Invest 91:709–716

Pichel JG, Lakso M, Westphal H (1993) Timing of SV40 oncogene activation by site-specific recombination determines subsequent tumor progression during murine lens development. Oncogene 8:3333–3342

Li, P, Tong, C, Mehrian-Shai R, Jia L, Wu N, Yan Y, Maxson RE, Schulze EN, Song H, Hsieh C-L, Pera MF, Ying Q-L (2008) Germline competent embryonic stem cells derived from rat blastocysts. Cell 135:1299–1310

Rémy S, Tesson L, Ménoret S, Usal C, Scharenberg AM, Anegon I (2010) Zinc-finger nucleases: a powerful tool for genetic engineering of animals. Transgenic Res 19:363–371

Robertson E, Bradley A, Kuehn M, Evans M (1986) Germline transmission of genes introduced into cultured pluripotential cells by retroviral vector. Nature 323:445–448

Rossant J, Nutter LM, Gertsenstein M (2011) Engineering the embryo. Proc Natl Acad Sci USA 108:7659–7660

Rueda N, Flórez J, Martínez-Cué C (2013) Apoptosis in Down's syndrome: lessons from studies of human and mouse models. Apoptosis 18:121–134

Ryan TM, Townes TM, Reilly MP, Asakura T, Palmiter RD, Brinster RL, Behringer RR (1990) Human sickle hemoglobin in transgenic mice. Science 247:566–568

Sauer B (1993) Manipulation of transgenes by site-specific recombination: use of Cre recombinase. Methods Enzymol 225:890–900

Schedl A, Beermann F, Thies E, Montoliu L, Kelsey G, Schutz G (1992) Transgenic mice generated by pronuclear injection of a yeast artificial chromosome. Nucleic Acids Res 20:3073–3077

Schedl A, Larin Z, Montoliu L, Thies E, Kelsey G, Lehrach H, Schutz G (1993) A method for the generation of YAC transgenic mice by pronuclear microinjection. Nucleic Acids Res 21:4783–4787

Schonig K, Bujard H (2003) Generating conditional mouse mutants via tetracycline-controlled gene expression. In: Hofker M, van Deursen J (eds) Transgenic mouse methods and protocols. Humana Press, Totowa, pp 69–104

Selfridge J, Pow AM, McWhir J, Magin TM, Melton DW (1992) Gene targeting using a mouse HPRT minigene/HPRT-deficient embryonic stem cell system: inactivation of the mouse ERCC-1 gene. Somat Cell Mol Genet 18:325–336

Simon-Chazottes D, Tutois S, Kuehn M, Evans M, Bourgade F, Cook S, Davisson MT, Guénet JL (2006) Mutations in the gene encoding the low-density lipoprotein receptor LRP4 cause abnormal limb development in the mouse. Genomics 87:673–677

Simon-Chazottes D, Frenkiel MP, Montagutelli X, Guénet JL, Desprès P, Panthier JJ (2011) Transgenic expression of full-length 2′, 5′-oligoadenylate synthetase 1b confers to BALB/c mice resistance against West Nile virus-induced encephalitis. Virology 417:147–153

Scherbik SV, Kluetzman K, Perelygin AA, Brinton MA (2007) Knock-in of the Oas1b(r) allele into a flavivirus-induced disease susceptible mouse generates the resistant phenotype. Virology 368:232–237

Skarnes WC, Auerbach BA, Joyner AL (1992) A gene trap approach in mouse embryonic stem cells: the lacZ reported is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. Genes Dev 6:903–918

Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, Mujica AO, Thomas M, Harrow J, Cox T, Jackson D, Severin J, Biggs P, Fu J, Nefedov M, de Jong PJ, Stewart AF, Bradley A (2011) A conditional knockout resource for the genome-wide study of mouse gene function. Nature 474:337–342

Smith DJ, Zhu Y, Zhang J, Cheng JF, Rubin EM (1995) Construction of a panel of transgenic mice containing a contiguous 2-Mb set of YAC/P1 clones from human chromosome 21q22.2. Genomics 27:425–434

Smithies O, Gregg RG, Boggs SS, Koralewski MA, Kucherlapati RS (1985) Insertion of DNA sequences into the human chromosomal betaglobin locus by homologous recombination. Nature 317:230–234

Stacey A, Bateman J, Choi T, Mascara T, Cole W, Jaenisch R (1988) Perinatal lethal osteogenesis imperfecta in transgenic mice bearing an engineered mutant pro-alpha-1(I) collagen gene. Nature 332:131–136

Stacey A, Schnieke A, McWhir J, Cooper J, Colman A, Melton DW (1994) Use of double-replacement gene targeting to replace the murine alpha-lactalbumin gene with its human counterpart in embryonic stem cells and mice. Mol Cell Biol 14:1009–1016

Stanford WL, Cohn JB, Cordes SP (2001) Gene-trap mutagenesis: past, present and beyond. Nat Rev Gene 2:756–768

Stevens LC (1960) Embryonic potency of embryoid bodies derived from a transplantable testicular teratoma of the mouse. Dev Biol 2:285–297

Stryke D, Kawamoto M, Huang CC, Johns SJ, King LA, Harper CA, Meng EC, Lee RE, Yee A, L'Italien L, Chuang PT, Young SG, Skarnes WC, Babbitt PC, Ferrin TE (2003) BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. Nucleic Acids Res 31:278–281

Sung YH, Baek IJ, Kim DH, Jeon J, Lee J, Lee K, Jeong D, Kim JS, Lee HW (2013) Knockout mice created by TALEN-mediated gene targeting. Nat Biotechnol 31:23–24

Tesson L, Usal C, Ménoret S, Leung E, Niles BJ, Remy S, Santiago Y, Vincent AI, Meng X, Zhang L, Gregory PD, Anegon I, Cost GJ (2011) Knockout rats generated by embryo microinjection of TALENs. Nat Biotechnol 29:695–696

Thomas KR, Capecchi MR (1987) Site-directed mutagenesis by gene targeting in mouse embryo derived stem cells. Cell 51:503–512

Utomo AR, Nikitin AY, Lee WH (1999) Temporal, spatial, and cell type-specific control of Cre-mediated DNA recombination in transgenic mice. Nat Biotechnol 17:1091–1096

Valancius V, Smithies O (1991) Testing an 'in-out' targeting procedure for making subtle genomic modifications in mouse embryonic stem cells. Mol Cell Biol 11:1402–1408

Van Keuren ML, Gavrilina GB, Filipiak WE, Zeidler MG, Saunders TL (2009) Generating transgenic mice from bacterial artificial chromosomes: transgenesis efficiency, integration and expression outcomes. Transgenic Res 18:769–785

Vitale-Cross L, Amornphimoltham P, Fisher G, Molinolo AA, Gutkind JS (2004) Conditional expression of K-ras in an epithelial compartment that includes the stem cells is sufficient to promote squamous cell carcinogenesis. Cancer Res 64:8804–8807

Vivian JL, Chen Y, Yee D, Schneider E, Magnuson T (2002) An allelic series of mutations in Smad2 and Smad4 identified in a genotype-based screen of N-ethyl-N-nitrosourea-mutagenized mouse embryonic stem cells. Proc Natl Acad Sci USA 99:15542–15547

Wagner EF, Stewart TA, Mintz B (1981a) The human beta-globin gene and a functional viral thymidine kinase gene in developing mice. Proc Natl Acad Sci USA 78:5016–5020

Wagner TE, Hoppe PC, Jollick JD, Scholl DR, Hodinka RL, Gault JB (1981b) Microinjection of a rabbit beta-globin gene into zygotes and its subsequent expression in adult mice and their offspring. Proc Natl Acad Sci USA 78:6376–6380

Wahnschaffe U, Bitsch A, Kielhorn J, Mangelsdorf I (2005a) Mutagenicity testing with transgenic mice. Part I: Comparison with the mouse bone marrow micronucleus test. J Carcinog 4:3

Wahnschaffe U, Bitsch A, Kielhorn J, Mangelsdorf I (2005b) Mutagenicity testing with transgenic mice. Part II: Comparison with the mouse spot test. J Carcinog 4:4

Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153:910–918

Wei C, Liu J, Yu Z, Zhang B, Gao G, Jiao R (2013) TALEN or Cas9—rapid, efficient and specific choices for genome modifications. J Genet Genomics 40:281–289

Wiznerowicz M, Trono D (2005) Harnessing HIV for therapy, basic research and biotechnology. Trends Biotechnol 23:42–47

Wong EA, Capecchi MR (1986) Analysis of homologous recombination in cultured mammalian cells in transient expression and stable transformation assays. Somatic Cell Mol Gene 12:63–72

Woychik RP, Stewart TA, Davis LG, D'Eustachio P, Leder P (1985) An inherited limb deformity created by insertional mutagenesis in a transgenic mouse. Nature 318:36–40

Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R (2013) One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. Cell 154:1370–1379

Yu T, Li Z, Jia Z, Clapcote SJ, Liu C, Li S, Asrar S, Pao A, Chen R, Fan N, Carattini-Rivera S, Bechard AR, Spring S, Henkelman RM, Stoica G, Matsui S, Nowak NJ, Roder JC, Chen C, Bradley A, Yu YE (2010) A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. Hum Mol Genet 19:2780–2791

Zhang F, Wen Y, Guo X (2014) CRISPR/Cas9 for genome editing: progress, implications and challenges. Hum Mol Genet. Apr 7. [Epub ahead of print]

Zhang Y, Proença R, Maffei M, Barone M, Leopold L, Friedman JM (1994) Positional cloning of the mouse obese gene and its human homologue. Nature 372:425–432

Chapter 9
The Different Categories of Genetically Standardized Populations of Laboratory Mice

9.1 Introduction

At the beginning of the twentieth century when genetics began to emerge as an experimental science, laboratory mouse resources were extremely limited. Apart from a few coat color mutants, which were bred by fanciers as pets, the only animals available for experiments were "albino" mice. These mice were bred with no specific mating protocol and were, in most cases, genetically heterogeneous. At that time, and based on the experience of dog and horse breeders, inbreeding was a practice to be avoided by all possible means because it was thought to lead to a decline in vigor and ultimately to the extinction of the colony. In fact, it is no exaggeration to say that the only qualities that were required of these "albino" mice were prolificacy, robustness, and tameness.

Regardless of the conditions under which they were bred, these mice (as well as the albino rats) were suited for most of the experiments that were undertaken at that time. For example, they could be used for performing experimental infections with a variety of pathogens, for the evaluation of the biological effects of drugs or for experiments in physiology, nutrition, etc. However, the transplantation of tumor cells, which was undertaken to study the process of cancer origin and progression, resulted in a high percentage of rejection, yielding unreliable results. Miss Maude Slye, for example, working at the University of Chicago, completed an extensive study concerning cancer inheritance in mice, but her contribution was nearly forgotten because she could not repeat her observations with "albino mice" from another supplier (Strong 1978).

At the same time, researchers began to develop a number of colonies with unique characteristics. At the origin of these colonies were a handful of mice selected for some interesting phenotypic traits. They were bred as closed populations, with no contribution from external breeders, to avoid losing the characteristics in question. Almost simultaneously, and from an increasing number of experiments involving tumor transplantations, it became progressively clear that the rate of success was higher when grafts were performed among mice of the same "family" than when grafts were performed among mice of unrelated origins.

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_9

In other words, instead of being disadvantageous, consanguinity appeared to be advantageous in the case of tissue/tumor transplantations. From this observation, as well as from a few others, it was then considered worth developing true inbred strains by systematically mating brothers to their sisters, generation after generation; the concept of inbred strains was born.

C.C. Little was the first to attempt the development of "pure" mouse lines by inbreeding (simultaneously, Helen D. King undertook the same sort of experiment with rats at the Wistar Institute). The first mouse inbred strain, *dba*, was started in 1909 by inbreeding mice homozygous for three recessive coat color alleles of independent origin (*d*: dilute, now *Myo5a*^{d}-Chr 9; *b*: brown, now *Tyrp1*^{b}-Chr 4; and *a*: non-agouti-Chr 2). Similarly, strain C57BL/6 was established in 1921, also by C. C. Little, from a cross between two "black" mice, female 57 and male 52, obtained from Miss Abbie Lathrop, a mouse supplier from Massachusetts. A few other strains were developed at about the same time by other scientists, in particular L.C. Strong in Cold Spring Harbor and N. Dobrovolskaia Zavadskaia in Paris.

Readers who are interested in the history of mouse genetics are invited to read the excellent book edited by Morse III (1978), which contains several chapters written by mouse geneticists of the early days, including L.C. Strong himself. In his chapter, Strong reveals the recipes he used for the successful development of his lines, and it is interesting to note that, a century later, most of these recommendations are still relevant. Some points from Strong's contribution are amusing anecdotes; for example, when he explains that he captured wild mice in a pigeon coop close to his lab, for "*sorting out their hereditary traits*" and was obliged, for practical reasons, to breed these mice "*under his bed in his honeymoon residence!*" These wild mice were also used by Strong to introduce some "vigor" in one of his stocks after an outbreak of "*paratyphoid*" (salmonellosis). Finally, and even though this is not clearly stated, it is more than likely that the wild-type allele at the agouti (*A*) locus, which is now homozygous in strains CBA and C3H (two of Strong's strains), is inherited from the wild mice trapped in the pigeon coop at Cold Spring Harbor on Long Island! Another interesting book on the early years of genetically standardized mice is *Making Mice* by Karen Rader (2004).

In addition to the contribution of North American researchers, which has clearly been fundamental for the development of most inbred strains of mice, one must also mention the contribution of Japanese scientists who established a number of colonies from fancy mice. Details of this contribution are reported in a book by Moriwaki et al. (1994).

Nowadays, a great variety of mouse strains are used routinely and many experiments involving laboratory mice would not be possible if these lines had not been patiently and carefully developed one century ago. In this matter, one must note that the mouse is, by far, the mammalian species with the largest variety of genetically defined strains, while the rat comes second, but far behind. In this chapter we will describe the genetic structure of the different types of genetically standardized strains and their optimal use in experimental genetics.

9.2 Inbred Strains

According to Hans Grüneberg, the "*introduction of inbred strains into biology is probably comparable in importance with that of the analytical balance into chemistry*" (Grüneberg 1952). This statement may have been considered somewhat peremptory 60 years ago but it does not appear exaggerated in the present context, as it has been validated by numerous examples. Indeed, and as we will discuss, inbred strains can be regarded as a "basic ingredient" in most experimental projects in mammalian genetics.

9.2.1 Inbred Mice are Isogenic and Homozygous at All Loci

According to the definition by the International Committee on Standardized Nomenclature for mice, "Strains can be termed *inbred* if they have been mated brother × sister for 20 or more consecutive generations, and individuals of the strain can be traced to a single ancestral pair at the 20th or subsequent generation". At this point the individuals' genomes will, on average, have only 1 % residual heterozygosity and can be regarded for most purposes as genetically identical (Fig. 9.1).

In practice, most of the mouse strains that are commonly used in research laboratories nowadays have undergone several tens of generations of brother × sister matings (indicated with an "F", for filial), and some of the most ancient (DBA/2 for example) may have passed 200 generations.

The definition of an inbred strain calls for a few comments. As stated, mice of the same inbred strain are genetically identical except, of course, for the sex-linked characteristics. One must note in particular that, because of strict inbreeding, all the mice of a given strain have become *homozygous* at all loci that were segregating in the founder ancestors (the original or ancestral breeding pair) and they all have become homozygous for the same allele (meaning that the maternal and paternal chromosomes are identical). This is also known as *autozygosity* because the two alleles are copies of the same ancestral allele. To describe this important characteristic, geneticists say that the mice in question are *isogenic*; in other words: they are genetically identical.

The process leading to homozygosity by progressive allele loss (or fixation) is easy to understand if we consider that, when by chance an allele that was present at generation F_n is not transmitted to at least one member of the breeding pair at generation F_{n+1}, it is then permanently lost. In other words, as inbreeding progresses, alleles are constantly lost but none are ever introduced, with the exception of rare de novo mutations. This, obviously, leads to both homozygosity and isogenicity!

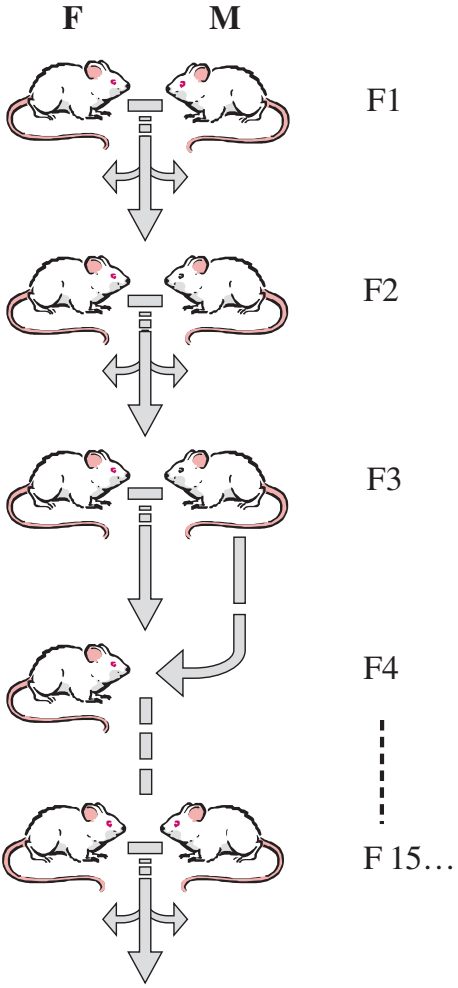


Fig. 9.1 *Inbred Strains*. This drawing represents schematically the breeding protocol commonly used to produce an inbred strain: mating a male and a female from the same litter (brother × sister) in successive generations. Theoretical calculation would indicate that parent × offspring exceptional matings (F4 in the example) would not affect the progression toward homozygosity provided that the parent selected for mating is the youngest of the pair. The uppercase letter F followed by a number represents the number of inbreeding generations. When this number is not known, a question mark is used: F? + 27, for example, would indicate that the number of brother × sister matings was not known when the strain was imported, but 27 generations of unrelaxed inbreeding have been added since this time. F13 + F28 indicates that 13 generations of strict inbreeding have been achieved in a breeding laboratory and an additional 28 in another laboratory

The categorization of the alleles that are lost or retained at each generation depends on chance for a large part, and if the inbreeding protocol could be reset with the same founder animals, it would lead to a strain with a different genetic constitution after the same 20 generations. This means that an inbred strain represents a unique and fortuitous assortment of alleles.

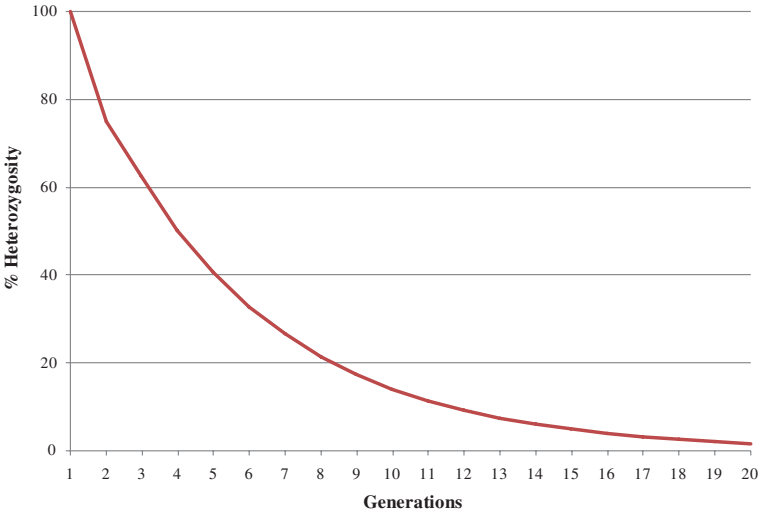


Fig. 9.2 *Effects of inbreeding*. The *curve* was drawn based on the ratios 1/1, 2/2, 3/4, 5/8, 8/16, 13/32, 21/64, and so on. In these ratios, the denominator doubles at each generation while the numerator is given by the Fibonacci sequence; each number being the sum of the two preceding numbers. This recursion relationship represents relatively accurately the decreasing percentage of genes that are still in the heterozygous state as inbreeding progresses. From generation F5 onwards, this percentage corresponds to ~19.6 % at each generation

To get a fairly accurate idea of what the genetic makeup of the individuals of an inbred strain actually looks like, one could imagine a totally virtual and theoretical experiment where the male pronucleus is removed immediately after fertilization, before it merges with the female pronucleus, while the remaining female pronucleus is duplicated, for example, after a short treatment with the alkaloid cytochalasin D, to become the diploid nucleus of a one-cell stage embryo. This totally artificial embryo would be a female, with the two chromosomes of each pair absolutely identical: this is precisely what the genome of all members of an inbred strain look like with the exception, of course, of the sex chromosomes.[1]

During the process of inbreeding, the progression toward homozygosity is relatively fast during the first few generations, where a great number of genes become homozygous, then it slows down and after 20 generations no more than 1–2 % of the loci that were heterozygous in the ancestors are still segregating. A mathematical series, based on Fibonacci's numbers, is traditionally used to model the decrease in heterozygosity as the number of sib-matings increases. Even though this curve is only an approximation, it represents fairly accurately the evolution of heterozygosity over time (Fig. 9.2).

When explaining the progression toward full homozygosity during inbreeding, we often consider the genome as a little bag full of genes, themselves considered

[1] In Chap. 6 we explained that, as a consequence of epigenetic modifications at the genome level, such a uniparental mouse could not exist in practice.

as independent entities. In reality, one must keep in mind that the genes are linked and arranged linearly on the chromosomes, and the evolution towards homozygosity involves blocks or "chunks" of chromosomes of variable sizes rather than individual genes. This explains why independent inbred strains carrying the same allele at a given locus have a great chance of sharing the same short segment of neighboring DNA (haplotype) on both sides of the allele in question, and this for historical reasons. For example, four strains homozygous for the albino (Tyr^{c}) allele (A, AKR, BALB/c, and SJL) are probably homozygous for the same short segment of chromosome 7 flanking the albino mutation (Tyr^{c}) because the mutation shared by these strains results from the same mutational event that occurred well before the creation of these strains (i.e., identical by descent or IBD). This peculiarity must be kept in mind because it applies to many other situations and may be advantageous (or unfavorable?) in the design of an experimental protocol. We will come back to this point in the section concerning congenic strains.

In most mammalian species, inbreeding of a natural population often has deleterious effects of variable intensity and phenotypic expression. In some (rare) cases, stillbirths are observed or newborns exhibit growth retardation and finally die. In other instances, there is a decrease in fitness or/and fertility, which is sometimes severe to the point that it leads to the extinction of the strain. All these adverse manifestations are commonly referred to as *inbreeding depression*. The basis of inbreeding depression has been debated over the last century, including by Darwin himself. Modern genetic studies suggest that inbreeding depression is predominantly caused by the presence of recessive deleterious mutations in natural populations that are progressively fixed in the homozygous state as inbreeding progresses (Charlesworth and Willis 2009). Alternative explanations, such as epistatic interactions, are also possible. In the mouse, surprisingly, inbreeding depression is not a serious issue as long as the breeders stem from the same natural population of closely related individuals. This is probably explained by the fact that wild mice, trapped in the same natural area, already have a relatively high percentage of consanguinity.

9.2.2 Inbred Mice are Genetically Stable in the Long Term

Around 230 different inbred strains were listed in the reference book by Michael Festing (1979). In 1998, 426 strains, with a brief description for each of them, were listed on the MGI website (http://www.informatics.jax.org/external/festing/mouse/STRAINS.shtml), but it is more than likely that many of these strains have been lost or terminated. However, among this impressive collection, two dozen have become very popular.

The fact that all members of the same inbred strain are genetically identical (isogenic) is certainly the major reason why they have become so prevalent in biomedical research. Scientists working with the same inbred strain, but in different laboratories or at different time periods, can perform experiments where the variations in the experimental results, by definition, will not be the consequence

of possible differences in the genetic constitution of the animals. In Chap. 10, devoted to the analysis of complex trait inheritance, we will provide examples showing that this is indeed a huge advantage.

Being isogenic also provides the great advantage that one can define, in detail and comprehensively, the phenotypic characteristics of each inbred strain by gathering phenotypic data concerning this strain from several laboratories and storing them in the same database. For example, The Jackson Laboratory has developed a program to collect baseline phenotypic data on the most popular inbred strains of mice through a coordinated international effort. Information collected in this program (*The Mouse Phenome Database*) is freely available to the community through the Internet (http://phenome.jax.org/) (Paigen and Eppig 2000). The establishment (and updating) of this database was made possible only because inbred mice are isogenic and accordingly genetically stable in the long term (Table 9.1).

Taking a look at the descriptions and genetic profiles of inbred strains in this database is always of great help when designing an experimental protocol. It may also contribute to saving animal lives by avoiding the repetition of experiments whose results are already known from experiments performed previously or elsewhere (for example, the dosage of a particular metabolite or the evaluation of a specific biological parameter).

Finally, being isogenic, mice of the same inbred strain are also *histocompatible* (or syngeneic). This means that they permanently accept tissue transplantations from any mouse of the same strain (and sex). Immunogeneticists have extensively used this peculiarity, since it allows studying the fate of cells with an immunological function in different contexts (cellular cooperation). It has also been extensively used (and still is) for the serial transplantation of malignant cells.

9.2.3 The Genetic Purity of Inbred Strains Must be Regularly Monitored

Although considered relatively stable in the long term, the genetic profile of a given inbred strain may change for two main reasons. The first results from accidental contamination by another strain; the second results from the progressive and insidious accumulation of novel mutations.

Genetic contamination resulting from the accidental mating of individuals of one inbred strain with another strain is by far the most important cause of alteration of the genetic profile. Such contaminations always result in a sudden and massive exchange of alleles and generally occur between strains that have the same or similar coat color (i.e., albino (Tyr^c/Tyr^c), agouti (*A/A*), or non-agouti (*a/a*)). These accidental crosses also occur between interstrain hybrid F1s and one of the parental strains, and between inbred and outbred strains (albino in particular). As a rule, accidental crosses result in an abrupt increase in breeding performances of the colony; such a change must always be considered suspicious and suggestive of a genetic contamination!

Table 9.1 This is part of a table listing in the Mouse Phenome Database

♀ Females	Mean	SD	*N*		♂ Males	Mean	SD	*N*	
129S1/SvlmJ	10.4	+0.494	$N = 8$		129S1/SvlmJ	10.6	+0.481	$N = 7$	
A/J	9.85	+0.495	$N = 8$		A/J	10.0	+0.489	$N = 8$	
AKR/J	10.2	±0.470	$N = 5$		AKR/J	9.36	+0.495	$N = 8$	▼
ALR/LtJ	8.82	±0.495	$N = 8$	▼▼	ALR/LtJ	8.75	±0.484	$N = 7$	▼
ALS/LtJ	10.1	±0.463	$N = 8$		ALS/LtJ	9.81	+0.449	$N = 7$	
BALB/cByJ	10.3	+0.515	$N = 8$		BALB/cByJ	10.0	+0.489	$N = 6$	
BPH/2 J	9.98	+0.471	$N = 7$		BPH/2 J	10.0	+0.445	$N = 7$	
BPL/1 J	11.7	+0.439	$N = 6$	▲▲	BPL/1 J	11.3	+0.438	$N = 7$	▲
BPN/3 J	10.8	±0.437	$N = 8$		BPN/3 J	10.5	+0.454	$N = 7$	
BTBR/J	8.85	±0.546	$N = 8$	▼▼	BTBR/J	9.82	±0.514	$N = 6$	
BUB/BnJ	9.58	±0.505	$N = 7$		BUB/BnJ	8.72	+0.480	$N = 4$	▼
C3H/HeJ	9.25	+0.495	$N = 7$	▼	C3H/HeJ	9.79	+0.470	$N = 7$	
C57BL/6 J	10.2	+0.584	$N = 114$		C57BL/6 J	10.3	+0.548	$N = 88$	
C57BR/cdJ	10.7	+0.439	$N = 7$		C57BR/cdJ	10.7	+0.443	$N = 8$	
C57L/J	10.7	±0.481	$N = 4$		C57L/J	10.8	±0.533	$N = 7$	
PWK/PhJ	10.3	±0.482	$N = 7$		PWK/PhJ	10.3	±0.493	$N = 8$	
RBA/DnJ	10.7	+0.444	$N = 7$		RBA/DnJ	10.6	+0.449	$N = 8$	
RBF/DnJ	9.78	±0.541	$N = 9$		RBF/DnJ	9.84	±0.525	$N = 8$	
RF/J	11.2	+0.454	$N = 7$	▲	RF/J	11.4	+0.447	$N = 8$	▲
Rill S/J	10.0	±0.438	$N = 8$		RIIIS/J	10.3	±0.432	$N = 7$	
SB/LeJ	8.63	±0.433	$N = 7$	▼▼	SB/LeJ	7.52	±0.437	$N = 8$	▼▼
SEA/GnJ	10.9	+0.438	$N = 6$	▲	SEA/GnJ	11.1	+0.436	$N = 7$	▲
SF/CamEiJ	9.06	+0.489	$N = 6$	▼	SF/CamEiJ	9.14	±0.463	$N = 8$	▼
SJL/J	9.84	+0.549	$N = 8$		SJL/J	10.6	+0.481	$N = 4$	
SKIVE/EiJ	10.1	±0.441	$N = 7$		SKIVE/EiJ	10.2	±0.449	$N = 8$	
SM/J	9.65	+0.546	$N = 8$		SM/J	9.62	±0.484	$N = 7$	
SOD1/EU	10.8	+0.495	$N = 8$		SOD1/EiJ	11.7	+0.464	$N = 7$	▲
SPRET/EiJ	10.8	±0.443	$N = 7$		SPRET/EiJ	11.6	±0.430	$N = 4$	▲

The red blood cell counts (measured in n/μL ± 1 SD) were performed on male and female mice of 72 different inbred strains (only 28 strains are represented here). Arrows indicate the most extreme values <http://phenome.jax.org/db/qp?rtn=views/measplot&brieflook=31802&projhint=CGDpheno1> Many other phenotypic parameters are stored in this Mouse Phenome Database for the most common inbred strains, allowing selection of the "best strain" when outlining an experimental project. Consultation of this database avoids wasting animals by repeating measurements uselessly

Mouse strains A2G and C57BL6/Ks are two well-known examples of inbred strains for which genetic contamination has been reported. A2G was considered to be a substrain of strain A until it was discovered that it probably originated from an "illegitimate" mating with an unknown partner. Mice of the A2G strain exhibit natural resistance to myxovirus (influenza), a peculiarity uncommon in most other laboratory strains, and it makes sense to believe that this characteristic is a "memory" of the illegitimate mating that occurred when the strain was developed. Strain C57BL/Ks (now C57BLKS) is another interesting case. The

strain derives from strain C57BL/6 but was contaminated with up to 25 % from the DBA/2 genome, 4 % from C57BL/10 J, from a 129 source and possibly some other undefined source. These untraced (and successive) contaminations were suspected for two reasons: because C57BL/Ks mice have a haplotype at the *H2* histocompatibility complex, which is not the one normally found in C57BL/6 mice (C57BLKS mice are $H2^d$, like strain DBA/2, instead of $H2^b$ like strain C57BL/6); and because congenic mice for the same obese (Lep^{ob}) mutation in these two backgrounds (C57BL/6J and C57BLKS) exhibited a different phenotype (Coleman and Hummel 1973). The suspicion of genetic contamination has now been molecularly documented and even cleverly used in an attempt to unravel the genetic causes of the background effect on Lep^{ob} phenotypic expression (Mao et al. 2006).

It is likely that many genetic contaminations have occurred in the past that have been rapidly detected and eliminated, but it is feared that the enormous increase in numbers of genetically engineered mouse (GEM) strains we are witnessing nowadays will exacerbate the threat of genetic contamination due to overcrowding of the breeding facilities. Commercial breeders are extremely sensitized to the risk linked with genetic contamination and perform regular monitoring of their stocks and strains. Most of them also have backups (archives) of their stocks cryopreserved in an embryo bank, allowing the rapid development of a fresh strain when necessary. At present, genetic monitoring of inbred strains is based on the use of molecular techniques at the DNA level and provides quick and highly reliable answers (See Box 9.1).

Box 9.1: Genetic monitoring of inbred strains and their derivatives

A variety of techniques has been described in the past to assay the genetic quality of inbred strains. All these techniques were based on the postulates that each inbred strain, as previously mentioned, is expected a priori to be homozygous at all loci. These techniques were designed following the progress of the genetic tools available for the species and consisted of analyzing a few traits, controlled by a set of specific alleles, and defining a specific pattern for each strain. Reciprocal skin grafting, for example, was extensively used in the 1960s because histocompatibility is controlled by many genes and requires complete genetic identity between the donor and the recipient. Skin grafting was a relatively inexpensive procedure, but unfortunately it was often influenced in both directions by environmental factors, yielding false-positive and false-negative results.

Analysis of the electrical charge of enzymatic proteins (isozymes) by electrophoresis in gels became popular in the mid-1970s because the technique was highly reliable and relatively easy to handle. However, this technique had the major drawback of being expensive to apply because each test required the use of specific and costly reagents.

Currently, most of the genetic monitoring techniques applied to inbred strains are based on DNA analysis and are extremely powerful. Most of

these tests are based on the analysis of strain-specific DNA sequences, revealed by routine techniques such as polymerase chain reaction (PCR) amplification of microsatellites (also known as simple sequence length polymorphism, SSLP), or SNP genotyping.

However, it must be kept in mind that the control of genetic purity should be undertaken in a broader context, considering parameters of very different nature (coat color, behavior, spontaneous diseases, breeding performance, etc.), and not only by applying sophisticated molecular techniques. Among these parameters, a careful observation of individuals from the same inbred strain, even if it may appear rather subjective, is always a very important source of information.

Genetic monitoring using microsatellite markers

Microsatellite markers are very popular because they are extremely easy to type at a very low cost (Benavides 1999; Mashimo et al. 2006). The technique consists of the amplification of short repeated sequences, in general dinucleotides of the type $(CA)_n$ or $(TA)_n$, with flanking primers using genomic DNA. There is an enormous number of microsatellite loci in the mouse and rat genomes (probably around 10^5), and it is generally not a problem to find a set of such molecular markers whose amplification products define a strain-specific pattern. This strain-specific pattern may be assayed on a sample of animals of the strain and compared to a reference DNA that is archived in the laboratory. Routine analysis of DNA samples with microsatellite markers will confirm isogenicity and, provided the markers have been carefully selected, it could also guarantee that the strain whose DNA is assayed indeed corresponds to its designation. One of the advantages of microsatellites is the fact that these are multiallelic markers, meaning that when tested in different inbred strains a marker will show several alleles (band sizes) for some of these strains. The use of fluorescently labeled primers for microsatellite loci combined with capillary electrophoresis represent a new, fast, and automated system for genetic monitoring (Mashimo et al. 2006). With this method, the resulting PCR products can be distinguished from one another by both their size and the fluorescent dye associated with them. The availability of different dyes allows the possibility of developing multiplex PCR (i.e., the combination of primers for multiple loci in one reaction) and pooling several PCR products into one capillary (Bryda and Riley 2008).

Genetic monitoring based on the use of single nucleotide polymorphisms (SNPs)

Genotyping for SNPs is an alternative approach that is now very popular for genetic quality control. SNP genotyping is inexpensive and can be performed in most research institutions or ordered to external companies. SNPs are the most common type of genetic variation observable at the DNA level and are found in both coding and non-coding regions. When localized in coding sequences, if the variant leads to an amino acid change or the creation of a stop codon, the SNP is said to be non-synonymous. On the other

hand, if the SNP does not change the protein sequence it is considered synonymous. Almost all SNPs are bi-allelic, presenting one of only two possible nucleotides (e.g., homozygous G/G or T/T) or both (e.g., heterozygous G/T) in an individual. Petkov and coworkers from The Jackson Laboratory (Maine, USA) have described the allelic distribution of 235 SNPs in 48 mouse strains and selected a panel of 28 such SNPs, enough to characterize most of the almost 300 inbred, wild-derived, congenic, consomic, and recombinant inbred strains maintained at The Jackson Laboratory (Petkov et al. 2004a). This set of markers encompassing all mouse chromosomes is an excellent tool for detecting genetic contaminations in mouse facilities by way of automated PCR systems. The same laboratory developed a new set of 1,638 informative SNPs selected from the publicly available databases and tested 102 inbred strains using Amplifluor genotyping (Myakishev et al. 2001). The selected SNPs are distributed approximately ~1.5 Mb apart across the mouse genome and, on average, 37 % will be polymorphic between any two inbred strains. Interestingly, these markers revealed subtle differences between closely related inbred strains and substrains, a result that was independently confirmed for the most popular C57BL/6 substrains: C57BL/6J from The Jackson Laboratory and C57BL/6N from the National Institutes of Health (Mekada et al. 2009; Zurita et al. 2011; Simon et al. 2013). SNP genotyping assays are currently based on allele-specific PCR (including KASPar fluorescent technology) (Nijman et al. 2008), real-time PCR (TaqMan®), direct sequencing, or DNA arrays (Moran et al. 2006). For those interested in the allele distribution of SNPs in different inbred strains, the Mouse Phenome Database presents the most comprehensive collection of SNPs, with more than 8 million unique loci and numerous inbred strains genotyped (see http://phenome.jax.org/db/q?rtn=docs/genonav).

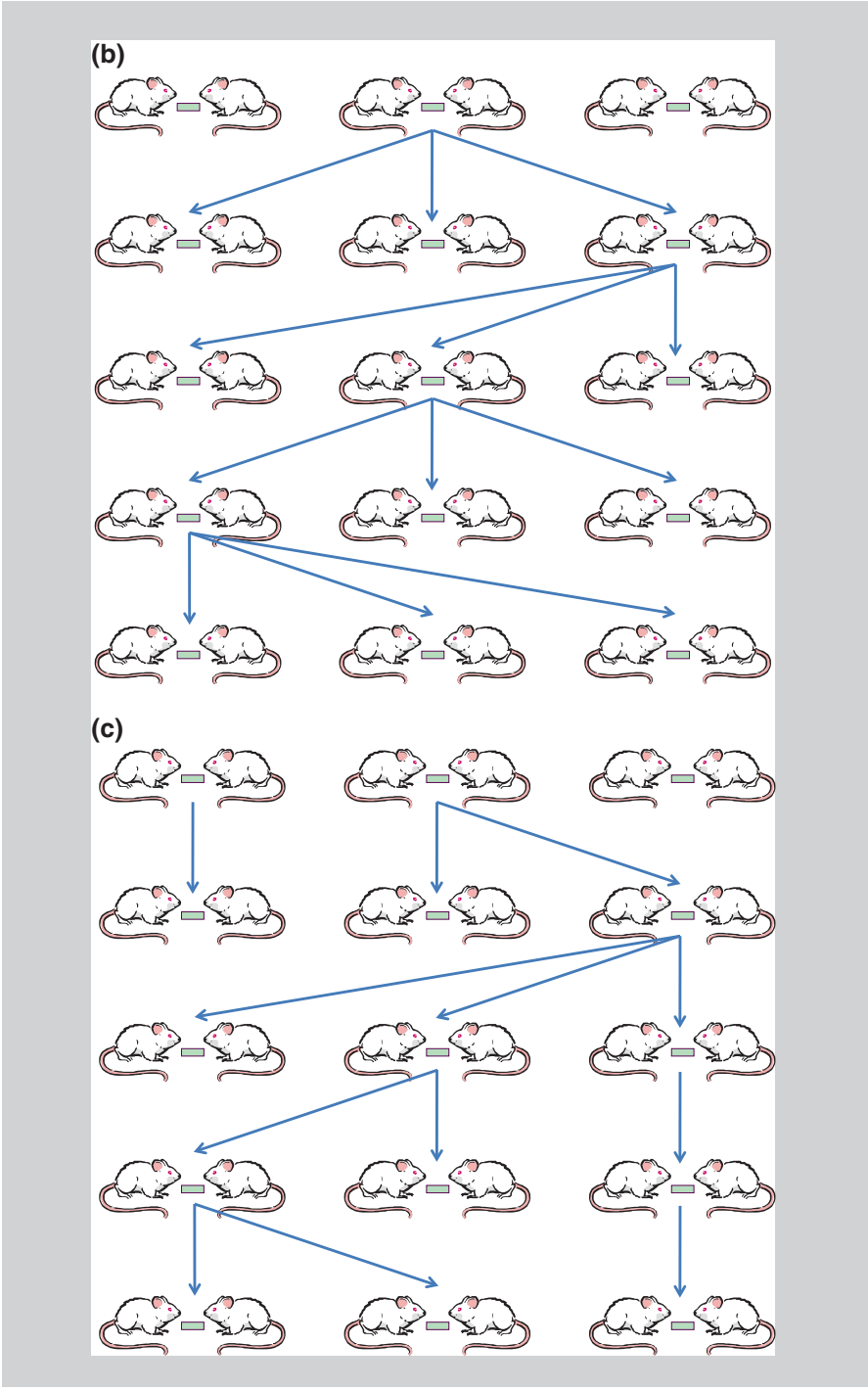
Box 9.2: Preserving the genetic purity of inbred strains

As discussed above, efficient techniques exist to monitor the genetic quality of inbred strains. However, once a strain is recognized as contaminated, the situation is irreversible. All individuals of the strain must be discarded, and another strain must be developed from a fresh set of breeders. The dramatic consequences of a genetic contamination imply that measures should be taken to prevent it. There are two efficient ways of preserving an inbred nucleus from genetic contamination: embryo freezing and complete physical isolation. Embryo freezing is, theoretically, the most efficient way of preservation because, once frozen, genomes are insensitive to mutations (DNA does not replicate), and of course contamination cannot occur. Sperm freezing may, in some circumstances, be used as an alternative to embryo freezing, but is of no use when a diploid genome must be preserved (Glenister and Thornton 2000).

Complete isolation of a breeding nucleus, for example, into a plastic isolator is a very efficient way of preserving genetic integrity and it is also an elegant way of preserving, at the same time, the health status of a rodent colony. Most of the commercial breeders of laboratory mice have chosen this strategy, which combines several advantages at a reasonable cost.

Box 9.3: Breeding protocols for the maintenance of an inbred strain

The system represented in **a** leads to the establishment of three (and not one) independent inbred strains, which are progressively divergent from one another due to genetic drift. In addition, if a strain stops breeding (a common situation in practice), it is then permanently lost. The system represented in **b** is certainly the best one, since, at each generation, three new pairs are established from generation N to breed mice of generation N+1. However, when the progenies are very small in size (a situation that is also common), it is not always possible to set the three new pairs of brothers and sisters. Finally, the system represented in **c** is the one that is generally used in practice.


(b)
(c)

Mutations are another source of genotypic change and are important to consider for two reasons: first, because their occurrence is completely beyond the control of the colony manager; and second, because they are insidious and in general impossible to detect by simple phenotypic observation or routine genetic monitoring. As reported in Chap. 7, the spontaneous mutation rates are quite low. They have been estimated to be in the range of 0.1 to 0.5 $\times$ 10^{-6} per locus per gamete for mutations towards a dominant allele and in the range of 0.6 to 0.8 $\times$ 10^{-6} per locus per gamete for mutations towards a recessive allele (Schlager and Dickie 1967). However, while a proportion of these new mutant alleles are effectively eliminated by inbreeding, another proportion may become progressively fixed in the homozygous state, replacing the original allele; this is one aspect of what geneticists call *genetic drift*. Genetic drift is a very slow and insidious process that is unavoidable. It contributes inexorably to strain divergence (and to the generation of substrains) when the same strain is propagated independently in different places.

Recently collected data concerning single nucleotide polymorphisms (SNPs) in different C57BL/6 substrains kept independently for a few years at The Jackson Laboratory indicated that the mutation rate for generating SNPs is very low (Wade et al. 2002). In addition, and assuming that only one SNP out of seven is translated into a functional polymorphism (see Chap. 7), this would suggest that the occurrence of new mutations is not a serious issue in the generation of sub-line divergence. The problem, however, is that the consequences of a novel mutation are not predictable. Mutations which are hidden in the genomes of substrains and can affect the outcome of an experiment are sometimes referred to as "*passenger mutations*" (Kenneth et al. 2012). There are many examples in the literature where substrains, although stemming from the same original inbred strains, have acquired new and unique phenotypic characteristics as a consequence of genetic drift (Bulfield et al. 1984; Stevens et al. 2007; Mattapallil et al. 2012). Mice of the C57BL/6J/OlaHsd substrain, for example, are homozygous for a deletion of the *Snca* locus (encoding for α-synuclein) on chromosome 6 (Specht and Schoepfer 2001). This deletion has modest phenotypic effects but might interfere in an unpredictable manner with other mutations if, for example, the C57BL/6J/OlaHsd substrain is used as a background strain for the production of knockout. In addition, a few spontaneous mutations have been reported to segregate differentially in the most popular substrains of C57BL/6 mice (C57BL/6J from The Jackson Laboratory and C57BL/6 N from the National Institutes of Health (separated in 1951), including a retinal degeneration mutation ($Crb1^{rd8}$) present in the N substrain and a deletion in the *Nnt* gene present only in the J substrain. The most comprehensive comparative phenotypic and genomic analysis of these popular strains has been recently published (Simon et al. 2013).

Similarly, if mice of substrain C3H/HeJ are experimentally infected with Gram-negative bacteria they may react very differently from mice of substrain C3H/OuJ. This is explained by the occurrence of a spontaneous mutation at the *Tlr4* locus (encoding for a Toll-like receptor) in the substrain C3H/HeJ, where all mice are homozygous for the defective allele $Tlr4^{Lps\text{-}d}$ (Poltorak et al. 1998). A very similar comment could be made for mice of the CBA/NJ substrain (CBA/CaHN-Btk^{xid}/J)

which, unlike mice of all other CBA substrains, are homozygous for an X-linked mutation (*Btk*^{xid}) producing a syndrome of immunodeficiency homologous to Bruton disease in humans (Berning et al. 1980).

What we have just said concerning the insidious and unavoidable occurrence of new mutations in an inbred strain also explains and justifies the recommendation by the International Committee on Standardized Genetic Nomenclature for Mice that inbreeding should never be relaxed. Inbreeding is inefficient in preventing mutations from occurring, but it contributes to the elimination of a substantial proportion of the new mutant alleles, and accordingly helps to preserve the genetic profile of a given strain in the long term. Similarly, the same international committee on nomenclature has decided that two strains with the same origin but separated in different colonies by 20 or more generations (for example, 12 in laboratory A and 10 in laboratory B) should be considered as two different substrains and designated appropriately (Davisson 1996; Wotjak 2003).

9.2.4 Most Inbred Strains are Derived from a Small Number of Ancestors

In 1982, two independent observations were published (Ferris et al. 1982; Yonekawa et al. 1982) reporting the remarkable structural homogeneity of mitochondrial DNA (mtDNA) among different strains of laboratory mice. This was quite surprising because the mitochondrial genome, a 16-kb double-stranded circular DNA molecule, was known in most species (including humans) to have a relatively high evolution rate. Considering that the mtDNA is inherited exclusively from the female parent, the most likely explanation to account for these observations is that all classical inbred strains share a common maternal lineage, probably inherited from a female albino mouse, bred as a pet, around 200 years ago. These observations, essentially based on the analysis of mtDNA restriction fragment length polymorphisms, have been recently confirmed and refined after sequencing and it has been confirmed that the laboratory strains are all derived from female progenitors of the *Mus musculus domesticus* subspecies, while *Mus musculus musculus* does not appear to have made any contribution to the mtDNA (Goios et al. 2007).

A few years after the observations concerning the mtDNA were published, similar experiments were made concerning the Y chromosome, a totally paternal contribution. Using five probes identifying Y chromosome-specific restriction fragments, six distinct Y chromosomes were identified among 39 standard inbred strains of mice, indicating that a minimum of six male mice contributed to the formation of the common inbred strains. Three Y chromosome types, distributed among 31 strains, were found to be of Asian (*M. m. musculus*) origin, while the remaining three Y chromosome types (8 strains) were of *M. m. domesticus* origin (Bishop et al. 1985; Tucker et al. 1992). All these findings are completely consistent with the historical records: they confirm that the laboratory inbred strains were

all derived from one or a few related females of the *Mus musculus domesticus* subspecies while the inter-strain polymorphisms represent the contribution of six males, all of them of the *Mus musculus musculus* subspecies.

9.2.5 Laboratory Inbred Strains have a Polyphyletic Origin

Inbred strains are often said to be artificial populations because their genetic constitution (isogenicity and homozygosity) has no natural equivalent. In fact, they could also be considered artificial populations because we now know, from historical records confirmed by extensive molecular data collected at the DNA level (sequencing), that they do not stem from one and a single subspecies of the *Mus* genus but from at least two: *Mus musculus domesticus* and *Mus musculus musculus* (Guénet and Bonhomme 2003). The finding of this polyphyletic origin was no real surprise if we recall the observations reported above concerning the origin of the mtDNA molecule and of the Y chromosome. This was also suspected for quite a long time, because it was the only way to explain that some electrophoretic variants of plasmatic proteins (for example, the esterase-2 allele *c* ($Es2^c$) or the phosphoglucomutase 1 allele *b* ($Pgm1^b$)), which are frequently found in laboratory strains as well as in mice of the *M. m. musculus* subspecies, are extremely rare in the genome of wild mice of *M. m. domesticus* subspecies (Bonhomme 1986; Bonhomme et al. 1987). The polyphyletic origin was confirmed and substantiated further after the complete high-resolution sequencing of the genomes of a large panel of inbred strains (Waterston et al. 2002; Wade et al. 2002; Yalcin et al. 2004; Frazer et al. 2007; Yang et al. 2011). In short, one can say that the genomes of laboratory inbred strains are a mosaic of chromosomal regions with distinct subspecific origins (Fig. 9.3).

On average, and according to the most recent estimates, the genetic contributions of the different *Mus musculus* subspecies is as follows: *M. m. domesticus* 68 %, *M. m. musculus* 6 %, *M. m. castaneus* 3 %, and *M. m. molossinus* 10 %. The remaining 13 % of haplotypes are of unknown ancestral origin.

It is also important to note that the distribution of diversity is markedly non-random among the chromosomes, with large regions of extremely low diversity and hot spots of diversity (Frazer et al. 2007; Church et al. 2009; Yalcin et al. 2011). This observation is particularly interesting because it results in an increase in genetic polymorphisms, making each inbred strain different from the other, and much more different from each other than we would have expected if mutations and genetic drift were the only source of diversity. Studies on the genetic determinism of complex traits benefit from this unique situation, as will be discussed in Chap. 10.

9.2.6 Inbred Strains Recently Derived from Wild Specimens

Over the last 20 years a variety of strains, derived from small nuclei of wild specimens trapped in well-defined geographical regions and belonging to well-characterized taxonomic groups, have been established in various laboratories.

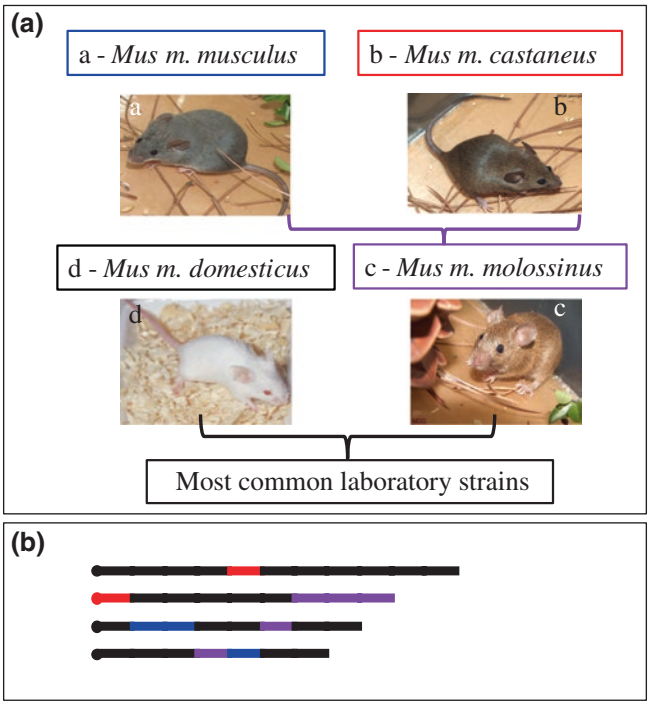


Fig. 9.3 *Origin of classical inbred strains of the laboratory mouse*. **a** Historical data, confirmed by sequence data, indicate that modern laboratory inbred strains derive from a small number of ancestors belonging to several different subspecies of the genus *Mus*. Today's classical laboratory inbred strains must be regarded as recombinant strains derived from four parental components (*in unequal percentages*): *M. m. domesticus, M. m. musculus*, *M. m. castaneus,* and *M. m. molossinus*. For this reason it would probably be more appropriate to designate them as *Mus* "*laboratorius"*! This polyphyletic origin explains (partially) the interstrain polymorphism segregating among the different laboratory strains. **b** The figure represents four mouse chromosomes in which some segments derive from one of the four ancestor subspecies (based on Frazer et al. 2007)

A list of these strains was published in the book *Genetic Variants and Strains of the Laboratory Mouse* (Bonhomme and Guénet 1996) and many of these strains are described on the internet at http://jaxmice.jax.org/list/cat481389.html. Most of these strains are now fully inbred with, in general, well over the required 20 generations of brother × sister matings.

Amongst all these inbred strains, special mention must be made of those derived from the *Mus spretus* species (for example, SEG/Pas, SPRET/Ei, and STF/Pas) because this species is one of those most distantly related to the laboratory strains (from the evolutionary point of view) that can still produce fertile hybrids. The production of these inter-specific hybrids results, in most instances, from natural matings between laboratory strain females and *Mus spretus* males, although some hybrids have also been produced with the opposite cross either by artificial insemination or by in vitro fertilization. F1 males with *Mus spretus* are sterile, as a consequence of the Haldane rule, but F1 females are fertile and can be used to produce

backcross progeny by mating them to males of either the laboratory strain or of the wild-derived strain. Segregating backcross progeny born to such F1 females, because of the very large number of allelic differences involved, have been instrumental for generating high-density/high-resolution genetic maps (see Chap. 4).

With the increasing use of techniques based on PCR amplification for the detection of genetic polymorphisms at the DNA level, the laboratory strains derived from *M. m. musculus*, *M. m. molossinus,* and *M. m. castaneus* have also been found to be of great value. Using a large set of oligonucleotides for the amplification of microsatellites, it has been demonstrated that 70 % of these primers yield PCR products polymorphic in length between strain PWK (an inbred strain derived from *M. m. musculus*) and the laboratory strain C57BL/6. By this criterion, the PWK strain appeared almost as distantly related as *Mus spretus* to the common laboratory strains (70 % vs. 74–84 %). With even more refined techniques, such as those based on the analysis of discrete structural variation that are capable of detecting a single base replacement (single strand conformation polymorphism or SSCP and denaturating gel gradient electrophoresis or DGGE), or even regional sequencing, it is likely that virtually any DNA stretch from a noncoding region that is more than 100 bp long would be found polymorphic between any inbred strain of wild origin and a reference laboratory strain. These observations should encourage scientists involved in gene mapping and/or positional cloning of genes to use the strains derived from *M. m. musculus*, *M. m. molossinus* or *M. m. castaneus* more intensely than those derived from *M. spretus*. They are easier to breed and frequently have the considerable advantage of allowing the production of F2 progeny where each individual results from two informative meioses and not just one—as is the case for the offspring of backcrosses.

The development and use of inbred strains recently derived from wild progenitors has been a great addition to the resources available to mouse geneticists and may prove even more useful in the future for the analysis of quantitative (complex) traits. However, one must be careful when using these new strains because the abundance of polymorphisms (SNPs or indels) sometimes makes difficult the establishment of causal correlation between the structural variations at the DNA level and their consequences in terms of gene function and expression. Examples of the advantages of using wild-derived inbred strains can be found in reviews on the subject (Guénet and Bonhomme 2003; Dejager et al. 2009).

9.2.7 Phylogenic Relationships Between Inbred Strains

Based on genotyping data collected by using a set of informative SNP markers and using an appropriate computer program for the optimal neighbor-joining method under the principle of maximum parsimony, a diagram has been established by researchers at The Jackson Laboratory (Petkov et al. 2004b), which represents the phylogenic relationships of the most commonly used inbred strains of the laboratory mouse (Fig. 9.4).

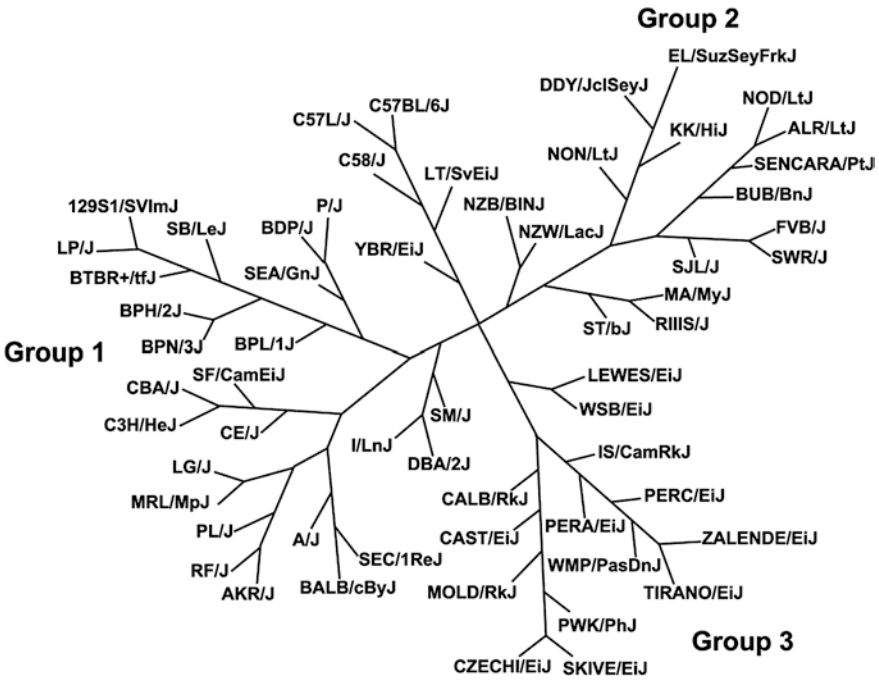


Fig. 9.4 *A mouse family tree*. The 60 inbred strains represented in this figure have been genotyped for a set of 1,465 informative SNP markers, evenly distributed over the whole genome (spaced on average <1.5 Mb). Applying the neighbor-joining method to the data, the authors constructed a family tree that could be organized into three groups: group 1, BALB/c, 129, and DBA-related strains; group 2, Swiss mice and Asian strains; group 3, wild-derived strains. The length and angle of the branches have been optimized for printing and do not reflect the actual evolutionary distances between strains. This family tree is in good agreement with most other existing genealogies (from Petkov et al. 2005). Using more markers for genotyping would increase the resolution of the phylogenetic tree (see, for example, Petkov et al. 2004b)

This diagram is in good agreement with the historical data previously collected (Beck et al. 2000) and can be used, for example, for the selection of closely or distantly related strains. This information is of primary importance for the design of an experimental protocol aiming to study the genetic determinism of inter-strain phenotypic differences. Indeed, selecting more distantly related parental strains when setting up a cross offers a greater chance of obtaining a higher resolution in the genetic analysis (Frazer et al. 2007).

9.3 Interstrain F1 Hybrids

Resulting from the cross of two inbred strains, F1 hybrids are heterozygous at all loci for which the parental strains have different alleles, but they are genetically uniform (isogenic) like their parents. Pairs of the same sex are equivalent to

monozygotic (identical) twins or to cloned mice. They are also histocompatible and permanently accept tissue transplantations from either parental strain, from their littermates, and from all their offspring; however, the parental strains will not accept a graft from the F1 hybrids.

F1 mice also exhibit the legendary hybrid vigor (heterosis), the opposite of inbreeding depression, making them the material of choice in many experimental protocols. This is common, for example, in the protocols aimed at the production of genetically engineered animals, where F1 hybrids are often used because of their high production of pre-implantation embryos that are highly resistant to manipulation (e.g., DNA pronuclear microinjection or for the creation of robust chimeras). However, a major drawback is that, when intercrossed, their progeny (F2) are genetically heterogeneous, since the alleles at all polymorphic loci start segregating due to recombination events in the F1 gametes during meiosis.

Interstrain hybrids can also be used to generate genetically heterogeneous populations. This is the case when, for example, F1 hybrids between strain A and strain B (abbreviated ABF1 or AXBF1) are crossed with F1 hybrids between strain C and strain D (CDF1 or CXDF1) to generate a four-way heterogeneous stock. In this case the basic ingredients of such a genetically heterogeneous stock (i.e., the original inbred strains A, B, C, and D) are perfectly identified, and similar, although not identical stocks can be produced at will when necessary. As we will explain later in this chapter, genetically heterogeneous stocks with an even more complex structure (for example, eight-way crosses stemming from eight different and unrelated inbred strains) have also been bred on a large scale for research in quantitative genetics (Threadgill and Churchill 2012) (see Chap. 10).

9.4 Co-isogenic and Congenic Strains

9.4.1 Co-isogenic Strains

When a mutation occurs in the breeding nucleus of an inbred strain, and if we assume that the new mutant allele has substituted the original one then the inbred strain in question differs from the original strain at one and only one specific locus. If the new mutation is viable and does not impair fertility, one can propagate the new strain by mating brother × sister mutant mice or, preferably, by mating, at each generation, a non-mutant mouse of the original inbred strain to an animal of the new mutant strain. These two strains are said to be *co-isogenic strains* or *segregating inbred strains* (Fig. 9.5).

Co-isogenic strains are extremely useful for gene annotation because they allow a comparison of the phenotypes of two allelic forms of a particular gene under optimal conditions (i.e., with no influence from the genetic background). A large number of such strains are stored worldwide in the major genetic repositories. Some inbred strains, like the famous C57BL/6, have several co-isogenic "companion" strains segregating for a variety of allelic forms involved, for example, in the determinism of

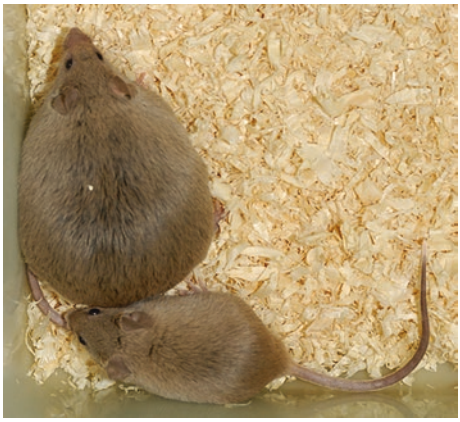


Fig. 9.5 *Co-isogenic strains*. The figure represents two mice of the same highly inbred strain DW/JPas. The obese mouse is homozygous for a short-sized duplication of the gene encoding the extracellular domain of the leptin receptor (*Lpr*). The ($Lpr^{db\text{-}Pas}$) mutant allele is inactive and the co-isogenic mouse grows to be obese

coat color. Mice of the C57BL/6-Tyr^{c} (albino) co-isogenic strain have become popular for the production of easily recognizable C57BL/6- +/+ ↔ C57BL/6-Tyr^{c}/Tyr^{c} chimeric mice from C75BL/6 ES cells injected into albino C57BL/6-Tyr^{c}/Tyr^{c} blastocysts (Schuster-Gossler et al. 2001).

Other strains, co-isogenic for mutations with detrimental effects on development or metabolism are also very interesting models because they can help in the analysis of pathophysiology, providing both the experimental animal and its control. Using such strains it is possible, for example, to attempt phenotypic rescues by grafting normal cells into a co-isogenic partner as a preliminary study for the design and development of possible therapies for human diseases. Co-isogenic strains, when developed in parallel to the background strain, may accumulate other genetic differences over time as a consequence of genetic drift. Thus, to minimize the effects of this drift, they must be periodically backcrossed to the original parental strain, or be cryopreserved.

Co-isogenic strains have two major drawbacks that are inherent in their origin and seriously limit their use: (i) they appear mainly as a consequence of a rare and fortuitous event (a mutation); and (ii), although they can appear in any inbred strain, it is in general not the strain that we would have been primarily interested in. For these two reasons, the use of co-isogenic strains is rather limited (see below).

9.4.2 Transgenic Strains are Equivalent but not Identical to Co-isogenic Strains

Genetically engineered mice can also be considered co-isogenic strains when the genetic modification is done in a way such that the targeted locus or transgene is the only difference from the wild-type animals. In the case of classical transgenic mice (additive or pronuclear transgenesis), this can be achieved by performing the pronuclear DNA (transgene) microinjection using embryos derived from an inbred strain

(e.g., FVB/N or C3H/He). Transgenic lines must be developed independently from each founder animal (microinjected embryos) and are normally kept by backcrossing the transgenic carriers (hemizygous Tg/0) with wild-type animals from the background strain and by selecting, at each generation, the new carriers (typically by PCR genotyping). One important difference between classical transgenic (by pronuclear microinjection) and co-isogenic strains is that the structure of the transgenic insertion can change with time: for example, in terms of copy number or DNA methylation. It can also be lost, leaving behind a micro-rearrangement, in general a micro-deletion.

9.4.3 Congenic Strains

Congenic strains are an alternative to co-isogenic strains with the advantage that any allele of the genome may be moved (geneticists would say "*introgressed*") into any inbred background. The disadvantage, as we will explain, is that the situation is not as pure, from the genetic point of view, as it is in the case of co-isogenics.

Congenic strains are produced by crossing two strains: the first one carries the allele or chromosome region of interest (i.e., spontaneous, induced or targeted mutations, as well as transgenes), and is referred to as the *donor strain*; the second strain is referred to as the *recipient strain* or *background* strain. The F1 offspring generated by crossing the above-mentioned two strains are backcrossed to the background strain, and the offspring that carry the allele of interest (i.e., the one originating from the donor strain) are crossed again to the background strain and so on, typically for ten or more successive generations.

During this succession of backcrosses, the chromosomes of the background strain progressively replace those of the donor strain, except for the one that carries the allele of interest. For this particular chromosome, the segment containing the selected or targeted allele is reduced in size only when a recombination event occurs that replaces a piece of chromosome of the donor strain with the homologous segment of the background strain. Since the occurrence of this sort of event depends upon the size of the segment, one then realizes that the chromosome carrying the targeted allele is gradually "eroded" on both sides, generation after generation, but in a nonlinear manner. The chromosomal segments flanking the selected locus have a tendency to remain associated with this locus, and this is the major difference between congenic and co-isogenic strains. In other words, while co-isogenic strains differ from the background strain at a single locus, congenic strains differ by a short chromosomal segment flanking the targeted locus, with the size of this segment being progressively reduced during the successive backcross generations.[2]

Since, on average, at each generation, an equivalent proportion of the background strain replaces one half of the genome of the donor strain, the progression of genome substitution is given by the formula $1/2^N$, where N is the number of backcross generations. This means that, theoretically, after 10 backcross

[2] The reduction in size of the introgressed chromosomal segment is in steps instead of linear.

generations only $1/2^{10}$ (<1/1,000) of the donor genome remains in the congenic strain. It is clear that this assumption is, again, purely statistical and the actual percentage of donor genome is subject to variations at each generation. In addition, and as we already pointed out, this estimation stands only for the chromosomes that do not carry the allele of interest (the *selected* or *targeted allele*). In the latter case, the reduction in size is a much slower process. According to Johnson (1981), if two loci *A* and *B* are distant by *c* Morgans, the probability that no recombination occurs between these two loci is e^{-c} per generation and, therefore, e^{-nc} after *n* generations. In the case of congenics, if *A* is the targeted locus and *B* a gene in the vicinity (located, for example, 10 cM from *A)*, the probability that the two loci remain in the same parental configuration after 10 generations is ~0.37 (=37 %). If *A* and *C* are 5 cM apart, the probability increases to 60.6 %, and it increases to ~90 % for two loci separated by 1 cM (0.01 Morgan). Stated differently, this means that there is only a 10 % (=100–90) chance that the segment harboring the introgressed gene will be smaller than 2cM (1 cM on each side) after a series of 10 backcrosses. This is not negligible since, as we discussed in Chap. 5, 1 cM of the mouse genome may contain up to 30–40 genes or even more, depending on the region (Fig. 9.6).

9.4.3.1 Marker-Assisted Congenics or Speed Congenics

The use of polymorphic and easy-to-score DNA markers has allowed a much more rapid and rigorous process of congenic strain development: the the so-called marker-assisted breeding (or backcrossing), also referred to as *speed congenics* methodology. The principle that underlies the speed congenics process is based on the fact that one can select the breeders, at each generation of backcrossing, based on the percentage of donor genome they have, by using either microsatellites or SNPs to distinguish the two parental strains. Obviously, the mouse with the lowest percentage of donor DNA is the one to select as a breeder for setting up the next backcross. Doing this greatly reduces the number of generations necessary to reach full congenicity (for example, from N10 to N5), and the strain development time, approximately by half.

At this point, it is important to note that, although a large number of molecular markers are necessary to perform efficient and reliable genotyping during the first backcross generation (in general 80–100 evenly distributed over the whole genetic map, for the N2 generation), this number decreases rapidly because, once a marker is typed "homozygous" for the allelic form of the background strain, it is no longer necessary to genotype the offspring of the future generations for this marker—it is permanently fixed (Markel et al. 1997; Wakeland et al. 1997). In order to fix the background Y chromosome, it is recommended to mate the female F1 hybrid to a male of the recipient strain early in the breeding scheme. Using molecular markers helps in the selection of breeders with the smallest amount of "flanking" or "hitchhiking" DNA, helping to alleviate the "*flanking gene*" concern (Wolfer et al. 2002; Chen et al. 2004). This requires the breeding of a large number of offspring, but these mice can be genotyped at an early age and discarded if considered unnecessary for future matings (Figs. 9.7 and 9.8).

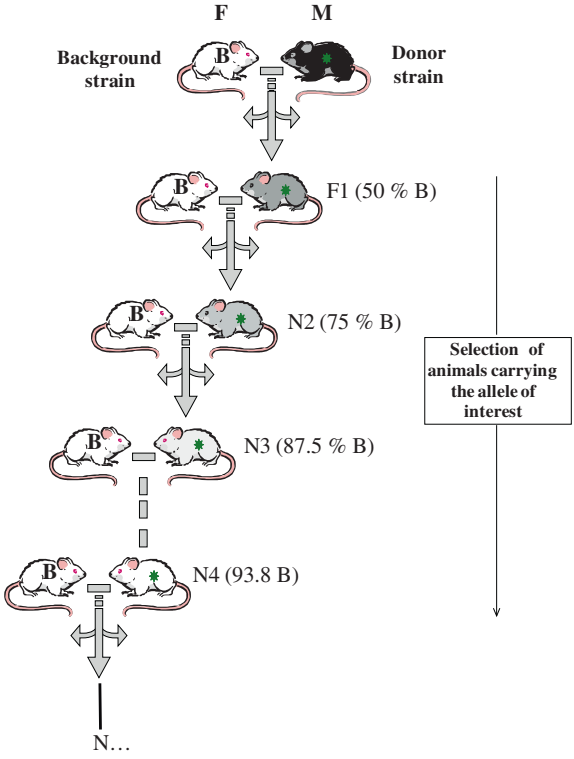


Fig. 9.6 *Congenic strains 1.* This scheme represents the successive steps in the establishment of a congenic strain. The initial step is a cross between two strains: a donor strain (*black* in the example) carrying the gene of interest (e.g., the targeted locus that can be a transgene or another allele) and a recipient or background strain (white in the example). At each generation, a breeder carrying the gene of interest (*) is backcrossed to a partner of the recipient (or background, B) strain. The degree of gray color indicates that, after each backcross generation, the offspring have an increased amount of the background genome. When the targeted gene has no easily recognizable phenotype, molecular genotyping is necessary. This genotyping is based on an easily detectable structural alteration (in most instances by PCR) within the locus in question. Closely linked markers may also be used

Everything described so far about how to establish a speed congenic strain corresponds to a standard protocol that can be applied in virtually any laboratory. In this strategy, the geneticist chooses the most "interesting" breeders for the intended purpose and mates them with an inbred partner of the background strain, then nature does the rest. In this context, the length of pregnancy and the time to reach sexual maturity are the only limits in the progress towards full congenicity. However, one can substantially accelerate the production of congenic strains by combining the efforts of geneticists and those of embryologists. One can choose, for example, 3-week-old females as heterozygous (carriers) breeders, superovulate them, collect their oocytes and perform in vitro fertilization with sperm from the background strain (as discussed in Chap. 2). The fertilized eggs (zygotes) can then be implanted into

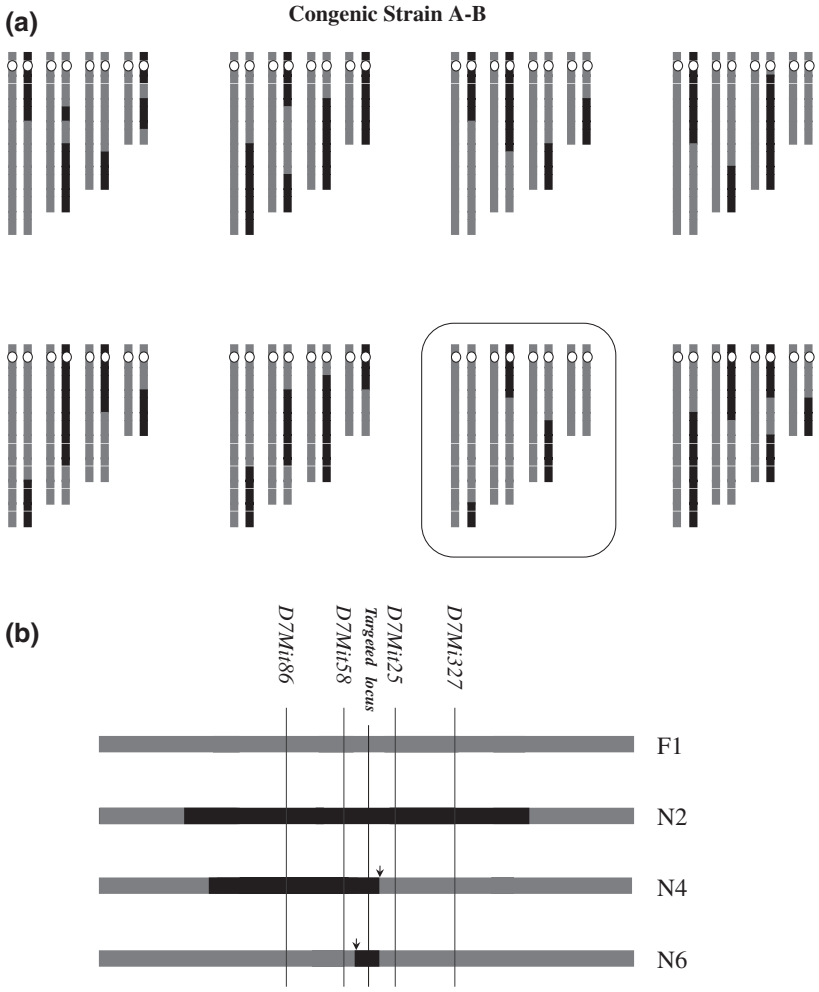


Fig. 9.7 *Congenic strains 2.* **a** After each backcross generation, 50 % of the genomic DNA of the donor strain (*black* chromosomes), on average, is replaced by the equivalent proportion of the genomic DNA of the background strain (*grey* chromosomes). With an appropriate genotyping assay, one can quantify the percentage of loci that are still heterozygous versus those that have become homozygous in the offspring of the backcrossed progeny (i.e., the mice that exhibit the lowest percentage of heterozygosity—boxed in the picture). Systematically selecting the breeders for the next (N + 1) generation among those with the lowest possible number of heterozygous loci is advantageous and speeds up the establishment of a congenic strain. The strategy can be used with any species and any markers. This is often called marker-assisted selection (MAS). **b** The chromosomal segments flanking the targeted allele are irrelevant and may generate difficulties in the interpretation of some experimental results. Genotyping with molecular markers allows the quality of the congenic strains to be increased by reducing the amount of irrelevant flanking DNA. For this, it is sufficient to retain as breeders the rare offspring with a recombination event between closely flanking markers and the targeted locus, as indicated in the figure. This selection can be perfectly applied after the two first backcross generations. A congenic strain with flanking regions of the "donor type" smaller than 1 cM is of top quality. The example shows microsatellite markers (polymorphic between the parental strains) flanking the gene (locus) of interest on chromosome 7

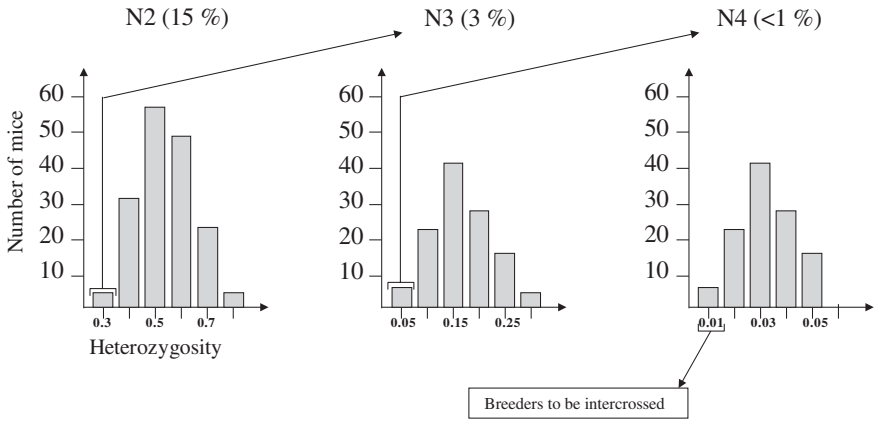


Fig. 9.8 *Speed congenics.* Selecting the breeder with the lowest percentage of introgressed (donor) DNA at each backcross generation requires the use of a great number of markers during the first generations of the breeding program. However, it is important to note that once a marker is typed "homozygous", it is no longer necessary to type it in the forthcoming generations. The bench work (genotyping) is then progressively reduced (from Wakeland et al. 1997)

pseudo-pregnant females and, when these females deliver their progeny, one can proceed with another round of selection with molecular markers. With an efficient protocol, the time to implement a new backcross generation can be reduced to 7–8 weeks, and a new congenic strain can then be established in no more than 10 months (*super-speed congenics*). In this regard, Japanese scientists have established a new record by injecting round spermatid nuclei from immature males (only 17 days old) into mature oocytes in vitro. With this technique called ROSI (for ROund Spermatid Injection), they were able to develop a full-congenic strain (N3 mice genotyped with 86 DNA markers) in only 106 days, a true *high-speed congenic* strategy (Ogonuki et al. 2009).

9.4.3.2 Congenic Strains and the Influence of Genetic Background

It is increasingly recognized that the genetic background (i.e., all genomic sequences other than the gene of interest) can influence the phenotype of an animal affected by a mutation. It has been shown that mutations (spontaneous and induced), transgenes, and targeted alleles (knock-outs and knock-ins) that are "moved" (introgressed) into a different background can exhibit a change in phenotype (Linder 2001; Doetschman 2009). This is mainly the result of the effect of several modifier genes. One of the first cases involved the classical diabetes (*Lepr*^{db}) mutation that presented transient diabetes in a C57BL/6 background but overt diabetes in C57BLKS (Hummel et al. 1972). Other examples include background effects on survival rate in *Egfr*$^{-}$ (epidermal growth factor receptor) knockout mice (Threadgill et al. 1995) and effects on tumor incidence and spectrum in *Trp53* and *Pten* knockout mice (Kuperwasser et al.

2000; Freeman et al. 2006), to name only a few. In order to avoid confounding or unreliable experimental results, particularly with the increasing number of mouse strains, attention to genetic background is crucial (Banbury 1997; Linder 2001).

A *Genetic Background Resource Manual* by The Jackson Laboratory is freely available at: https://secureweb.jax.org/jaxmice/literature/geneticBackground.html. This 12-page booklet contains a series of examples where the genetic background has been misleading and explains how to take this into account in experiments involving mice. We strongly recommend it.

9.4.3.3 Congenic Strains and the Genetic Determinism of Complex Traits

As we will discuss in Chap. 10, congenic strains have been extensively used since the early days of mouse genetics and still are. They are particularly suited for the genetic analysis of phenotypes that are controlled by several genes, and it is precisely by developing such strains that George D. Snell and his colleagues from The Jackson Laboratory could elucidate the genetic determinism of histocompatibility (Snell 1948).[3]

As we already mentioned at the beginning of the present chapter, tissue transplantations performed between mice belonging to unrelated populations—for example, mice from two different inbred strains—are rejected. On the other hand, the same transplantations performed between any two mice of the same inbred strains (and the same sex) are permanently accepted. The problem is that, in the case of tissue transplantations, the rejection, which is the observed phenotype, is controlled by several loci, each of them independently triggering the same phenotype. To clarify the situation, Snell bred a series of strains with the same C57BL/10 genetic background, but congenic for a single Mendelian unit inducing tissue incompatibility. To simplify the analysis of the phenotype and to save time, Snell injected tumor cells into mice segregating for the histocompatibility gene (all symbolized by *H*). At each generation, only the mice that survived were "selected" and accordingly were "resistant" to the (tumoral) tissue transplantation. He called these congenic mice *congenic-resistant* (CR) and developed a very clever protocol to characterize each of these strains, thus avoiding duplications (CR strains congenic for the same *H* locus just by chance). By doing this, Snell succeeded in making an inventory of many of the *H* loci segregating among the laboratory strains. This strategy could be adapted with almost no change to the genetic analysis of any trait that is under polygenic control; for example, resistance to infectious diseases. When a congenic strain has been established, there is still a lot of work to do to finally characterize the gene involved in the

[3] G.D. Snell, J. Dausset and B. Benacerraf were awarded the Nobel Prize in 1980 *"for their discoveries concerning genetically determined structures on the cell surface that regulate immunological reactions"*.

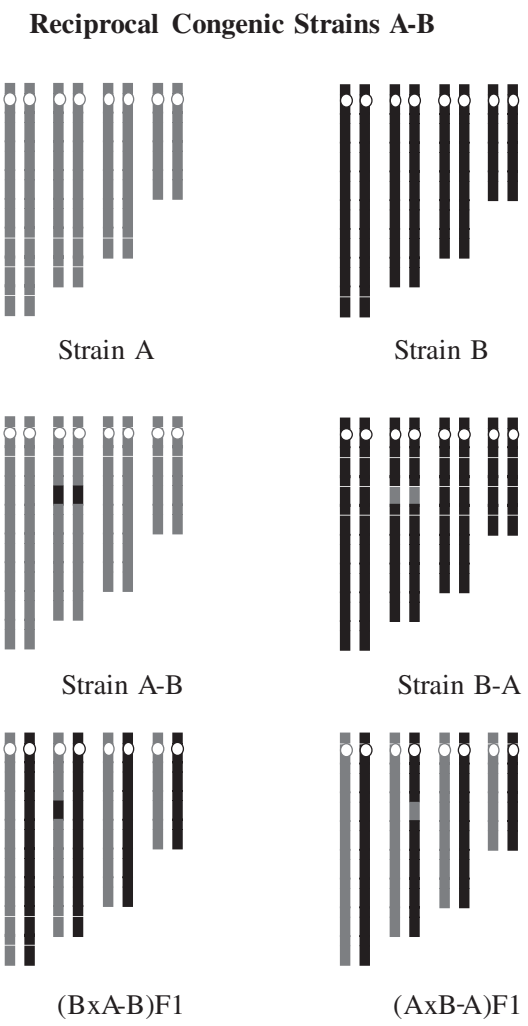


Fig. 9.9 *Reciprocal congenics.* Reciprocal congenic strains allow for comparisons to be made with a high degree of standardization because epistatic interactions may be controlled by this procedure. If, for example, a given allele in the background of a congenic strain interferes with the expression of the introgressed gene, this might be detected when comparing the two reciprocal congenic strains or the F1 of these congenic strains with the parental inbred strain, as indicated in the figure

phenotype, given that the chromosomal fragment is sometimes quite large. Nevertheless, congenic strains are certainly of great help in these investigations.

A few more comments are necessary to complete our description of the congenic strains. The first refers to the fact that a pair of congenic strains can be perfectly established even if the donor strain is not inbred. For example, if a mouse is identified with an interesting characteristic segregating in a non-inbred population, it is possible to derive one or more strains congenic for this trait, following the same protocol described above.

Another interesting possibility is to develop reciprocal congenic strains by introgressing a specific locus of strain A into the background strain B and, reciprocally, the homologous locus of strain B into the background strain A. At the end

of the experiment, one has a total of four strains: the two parental inbred strains A and B on the one hand, and the reciprocal congenic strains A_B and B_A on the other. One can then compare the F1 between strain A and the congenic strain B_A with the reciprocal F1 hybrid between strain B and the congenic strain A_B. This type of experiment, making use of F1, has the advantage of eliminating the side effects of possible epistatic interactions with the genetic background and is likely to provide more reliable answers (Fig. 9.9).

Finally, a comment is warranted on the use of congenic strains as tools for the analysis of quantitative (or complex) traits. When we discussed the experiments by Snell regarding the genetic analysis of histocompatibility, we mentioned that the derivation of CR strains made possible the individual identification of several *H* loci. Of course, this identification exclusively concerns the genes that are in a different allelic form in the congenic partners; those that are non-polymorphic remain undetected. This may appear to be a truism, but keeping in mind that the classical inbred strains of laboratory mice were all derived from a small pool of ancestral progenitors, it is clear that the experiments by Snell made possible the discovery of only a small proportion of all the *H* genes of the mouse species. Many other loci remained undetected, and it is likely that the derivation of new CR strains from wild mouse specimens would certainly be very rewarding. This comment applies, of course, to all situations where many genes (and many alleles) are involved in the determinism of a complex or quantitative trait (See Chap. 10).

9.5 Consomic Strains

Consomic strains, also called *chromosome substitution strains* (CSS), are a variation of the congenic strains concept in which the introgressed DNA is a complete chromosome, rather than a piece of chromosome flanking a given gene (Nadeau et al. 2000). These strains have been very useful for the rapid mapping of phenotypic traits to a specific chromosome. They are also useful for the detection of chromosomal regions (the so-called quantitative traits loci, QTLs) having an influence in the determinism of a particular phenotype (for example, the resistance to or susceptibility for carcinogenesis). This point will be explained in some detail in Chap. 10. Only a few sets of consomic mouse strains are available, but it is likely that other sets will be developed in the future to accompany the development of investigations in multifactorial inheritance (Gregorova et al. 2008; Mattson et al. 2008) (Fig. 9.10).

Using a marker-assisted protocol, consomic strains are easy to produce. However, one must keep in mind that tiny pieces of chromosomes of the donor strain might escape the marker-assisted selection process if, by chance, they are not identified by a marker. In the same way, there is no guarantee that the telomeric region of a given chromosome pair is transferred intact since there is, in most instances, no distal marker to check this. Finally, and according to the available information, attempts to develop a full set of inter-specific consomic mouse strains from distantly related mouse species or subspecies (for example, *Mus spretus* as a donor strain and

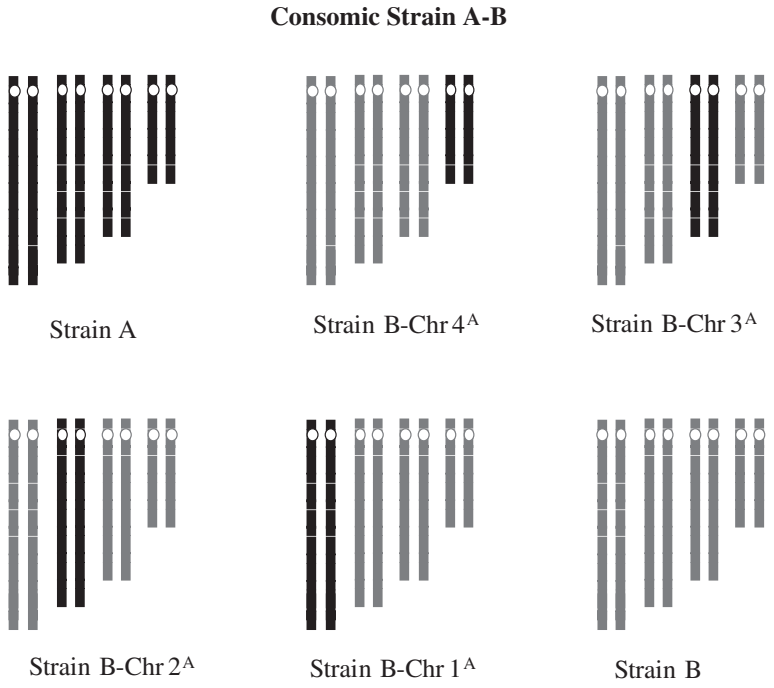


Fig. 9.10 *Consomic strains.* A consomic strain is an inbred strain in which one of the chromosome pairs has been replaced by the homologous chromosome pair of another inbred strain after a series of marker-assisted backcrosses. A complete panel of consomic strains consists of 21 strains, each derived from the same donor and host strains but having each a different chromosome pair (Chr 1–19, X or Y) of the host strain replaced by its homolog from the donor. A reciprocal panel can be produced by inverting the donor and host strains. One can never be sure that two strains are fully consomic for the telomeric ends because telomeric markers are often missing

C57BL/6 as a background strain) have proved difficult or even unsuccessful, presumably because of deleterious epistatic interactions between genes (or alleles) that have been separated by evolution for a long time (over 1.5 Myr). For example, some hybrid sterility genes have been reported that will result in complete sterility of the F1 male offspring between *Mus spretus* (any strain) and laboratory strains (Forejt 1996).

9.6 Recombinant Inbred Strains and Recombinant Congenic Strains

Recombinant inbred strains (RIS) are developed by crossing two parental inbred strains to generate F1 hybrids and then intercrossing these F1s to generate F2s. Finally, randomly chosen F2 animals are then brother × sister mated for 20 or more generations to develop a group of related inbred strains (Bailey 1971). RIS

are grouped by sets (also referred to as panels): a collection of RIS derived from the same parental strains. For example, the C57BL/6 × DBA/2 (BXD) is, at the moment, the largest mouse RI panel with ~90 strains. These are true inbred strains, meaning that they are homozygous at all loci but have the additional characteristic that each RIS has a unique fixed combination of the parental alleles in a 50:50 proportion (on average). For example, each strain of the set of 33 AXB-BXA strains, derived from the initial cross of a C57BL/6 mouse with a A/J mouse, carries either the B6 allele or the A allele at each locus of its genome; by typing all of these allelic forms, one can establish a strain distribution pattern (SDP) for each of the strains, which lists the collection of alleles inherited from either the parental strain A or the parental strain B6. Of course, this SDP is fixed forever in each strain (not taking into account the rare mutations that inevitably occur), and new data are constantly added to it, allowing correlations to be made between genotypes and phenotypes simply by scanning, generally with the help of a simple computer program, the co-segregation of a new phenotype (or genotype) with the existing SDP. RIS have proved very helpful when used for gene mapping, in particular for the rapid regional assignment of microsatellites on a given chromosome, when these markers were cloned by the thousands for the establishment of high-density genetic maps (see Chap. 4). They have also been used for the mapping of chromosomal regions (QTLs) involved in the genetic determinism of some behavioral characteristics (for example, taster/non-taster for a chemical compound, alcohol intake, etc.) or of some immunological responses, and they will very likely still be of great help in many other experiments where the phenotype is measured on a group of animals rather than on individuals (Zou et al. 2005) (Fig. 9.11).

Recombinant congenic strains (RCS) are similar to RIS in their genomic structure except that the proportion of the parental alleles in a given strain is not 50:50 but 75:25 or 87.5:12.5, depending on the set (Demant and Hart 1986). This is achieved by inbreeding mice of the first or second backcross generation to one of the parental inbred strains (the background strain). As we will explain in Chap. 10, RCS are helpful for identifying genes associated with polygenic inheritance, especially when the number of genes is high. RCS with a small percentage of introgressed genome in a background strain have a greater power of resolution, and their use increases the likelihood of zero or only one single locus governing the studied phenotype (QTL) being isolated in a given RCS. For example, RCS have been very helpful for unraveling the genetic determinism of colon cancer in the mouse (Demant 2003).

Interspecific recombinant congenic strains (IRCS) have also been developed from the parental strains C57BL/6JPas and SEG/Pas (*Mus spretus*) (Burgio et al. 2007). This set of strains has proved particularly useful for the analysis of the genetic determinism of some anatomical traits (Burgio et al. 2009). The differences between congenic strains and recombinant congenic strains is that, in the case of congenic strains, the introgresssed region(s) is unique, with the smallest possible size, and chosen a priori by the investigator, while there is in general more than one region in the case of RCS, with these regions being of variable size and not selected by the investigator. This being taken into account, and provided the strain combination is appropriate it is clear that it may sometimes be advantageous to choose a specific RCS as a donor strain for the development of a congenic strain.

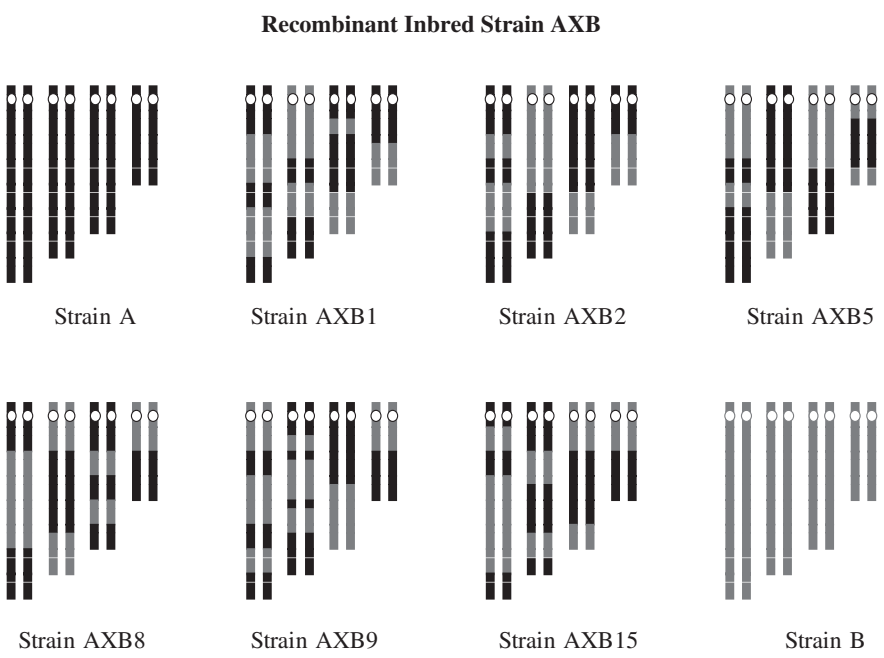


Fig. 9.11 *Recombinant inbred strains.* This diagram represents a panel of six recombinant inbred strains (RIS) flanked by the parental strains A and B. These strains derive from the same initial cross but each have a unique combination of loci derived by recombination of the alleles present in the original parental strains. Since RIS are inbred and each strain has a unique genotype, RIS have a number of advantages over F2 or backcross mouse populations as tools for mapping genes or quantitative trait loci (QTLs)

9.7 The Collaborative Cross

The panels of RIS described in the preceding section represent a first-rate resource for the identification and analysis of the genetic determinants of complex traits. Since all the mice within a given strain have the same genotype, phenotyping can be carried out on groups of varying sizes, yielding a phenotype that can be expressed in terms of percentage with a confidence interval that is only dependent on the size of the sample. Using RIS allows assessing the genetic determinism of susceptibility to certain drugs, to certain forms of cancers, and to experimental infections with pathogens. These types of experiments would be difficult, if not impossible, to achieve by the mere genetic analysis of F2 or backcross populations. Another advantage of the RIS is to reduce the cost of the experiments. Indeed, given that most of the existing strains are already genotyped for a large number of genetic markers it is in general easy to detect co-segregation of one or a few specific marker(s) with the data collected from phenotyping. Unfortunately, because they are all derived from a handful of classical inbred strains, the different panels of available RIS display a relatively low level of genetic diversity

when compared to the diversity found in the *Mus* genus as a whole and this often appears as a limitation in the use of RIS.

Considering these advantages and drawbacks in the use of the RIS panels stimulated discussions among a group of researchers interested in quantitative genetics (*The Complex Trait Consortium*), and these discussions led to the idea to develop a new resource, better adapted to the analysis of complex traits. Nowadays, this resource is being actively developed and it is known as *Collaborative Cross* (Fig. 9.12 a, b). The Collaborative Cross (CC) is an extension of the recombinant inbred strain concept with however a much higher power of resolution and a much higher level of genetic diversity (Churchill et al. 2004; Chesler et al. 2008; Threadgill et al. 2011). The Collaborative Cross is derived from a panel of eight carefully selected founder inbred strains that consist of: (i) three classical, traditional inbred strains (A/J, C57BL/6J, 129S1/SvImJ); (ii) two inbred strains affected by a genetically complex pathology (diabetes/obesity) NOD/LtJ, NZO); and (iii) three inbred strains derived from wild progenitors of the three main subspecies of the *Mus* genus (CAST/Ei derived from *Mus m. castaneus*; PWK/PhJ derived from *Mus m. musculus* and WSB/Ei derived from *Mus m. domesticus*). The eight founder strains were first crossed pairwise to generate all [= (8 x 7) / 2 = 28]

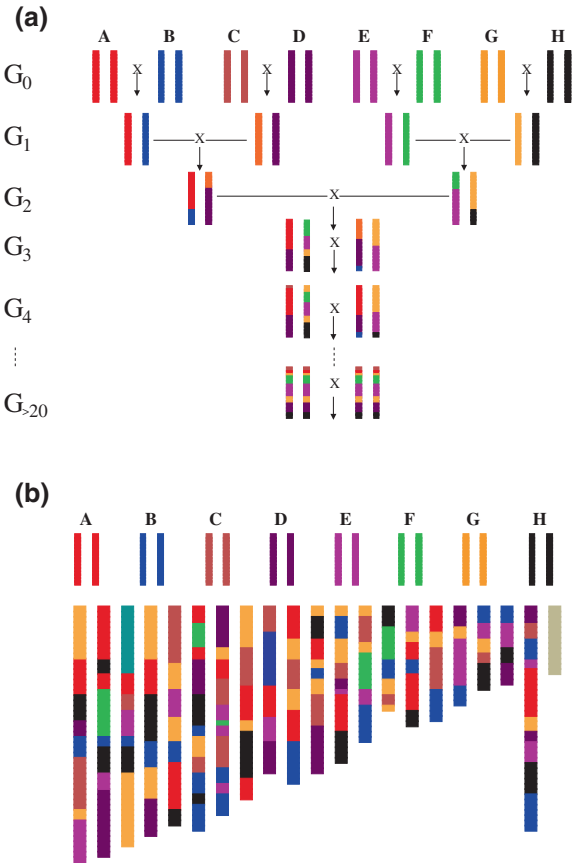


Fig. 9.12 *The Collaborative Cross.* **a** This is a randomized cross of eight unrelated mouse inbred strains selected by the members of the Complex Trait Consortium. The strains are first crossed pair-wise to make all ((8 x 7) / 2 = 28) possible G1. A set of possible four-way crosses is performed, keeping Y-chromosome and mitochondrial balance. Finally, all eight genomes are brought together in G2:F1, and the offspring of this cross are inbred. The Collaborative Cross is a community resource that was initially designed for the purpose of mapping complex traits. **b** The initial plans were to breed around 1,000 inbred strains where all the alleles of the initial inbred strains would be associated with a wide and unique variety of combinations. The illustration presents the 19 autosomal chromosomes plus X and Y of a hypothetical RIS from the Collaborative Cross

possible G1 hybrids, then a balanced subset of non-overlapping 8-way progeny was selected and brother x sister mated for several (i.e. ≥ 20) generations to produce a very large set of RI lines with variable proportions of the genomes of the parental strains. These RI lines will allow for the detection of biologically relevant correlations between thousands of traits with an unprecedented power of resolution. Such a panel of CC lines is currently being developed in three laboratories: the University of North Carolina in Chapel Hill, the Tel Aviv University, and the University of Western Australia in Perth. For most of these CC lines the progress of inbreeding is monitored by PCR genotyping of the SNPs segregating among the parental strains. Nowadays a few tens of lines, displaying > 90 % homozygosity, have already been made available to the community in particular to some "Mouse Clinics", for extensive phenotyping. For example, a comprehensive and comparative phenotypic analysis is under development at the German Mouse Clinic (Helmholtz Zentrum München-in Neuherberg-http://www.mouseclinic.de/), with more than 500 parameters being tested including parameters that characterize allergy, behavior, blood chemistry, bone and cartilage structure, energy metabolism, eye and vision, immunology, lung function, neurology, nociception, and pathology. An example of the power of the CC lines for the analysis of complex traits was reported in the case of experimental infections with the Ebola virus (Rasmussen et al. 2014). Unexpectedly, while none of the classical laboratory strains display the whole range of symptoms commonly associated with human Ebola haemorrhagic fever, after experimental infection, strains from the Collaborative Cross exhibited a variety of phenotypes ranging from complete resistance to a severe (lethal) haemorrhagic syndrome. These observations indicate that the genetic background strongly determines susceptibility of mice to Ebola hemorrhagic fever and this opens avenues for the development of a better animal model for the study of human infection.

At its conception the project of the Complex Trait Consortium, was to generate 1,000 RI lines recombinant for variable genomic proportions of the eight parental strains enabling detection of biologically relevant correlations between thousands of measured traits with an unprecedented power of resolution (135,000 recombination events!). Practical concerns related to the facilities needed to host all these strains, their distribution, their health status and the difficulties to raise funding to preserve them have led researchers to consider some reduction in the generation of new CC lines. At the end of 2014 around 100 strains were available for research purpose. These CC strains are all, at least, 90 % homozygous and derive from at least 6 and in many cases from all 8 founder strains. In parallel to the breeding of these strains a genotyping project involving more than 77,000 maximally informative SNPs is under development. This project, known as the MegaMUGA (for Mega Mouse Universal Genotyping Array), is to cover the genome of every one of the CC lines with a very high density of SNP markers (average spacing of 33 kb). Even if it may take some time before the Collaborative Cross project is completed, it is now a reality with multiple applications. For more information concerning Mega MUGA refer to: http://csbio.unc.edu/CCstatus/Media/MegaMUGAFlyer.pdf.

9.8 Outbred and Random-Bred Stocks

Outbred and random-bred stocks are populations of laboratory animals that are radically different from those we considered above in the sense that they are genetically heterogeneous, or *heterogenic* as we might say to keep the same sort of terminology. According to the official definition, outbred mouse stocks are "*closed populations (for at least four generations) of genetically variable animals that are bred to maintain maximum heterozygosity*". Compared with inbred strains, F1 hybrids, or congenic strains, the genetic constitution of a given animal, taken randomly from an outbred stock, is not known a priori and must be defined when necessary.

Outbred mice represent the bulk of laboratory animals sold by commercial vendors for the purpose of experimentation. These animals are usually bred according to a system that minimizes (or, more exactly, reduces) inbreeding, and accordingly contributes to the maintenance of a certain amount of heterozygosity in the population (Hartl 2001). A classical breeding scheme for these populations would consist, for example, of the mating in room C and D of n males originating from room A with the equivalent number of females taken from room B, with n being as great as possible. For the production of the next generation (G + 1), the breeding scheme would be similar with n males from room C being mated with n females of room D, and so on. Doing this, generation after generation, the polymorphic alleles that were segregating in the population at generation G have the greatest chance of still being represented at generation G + 1 in roughly the same proportion. The greater the samples of breeders used for the production of G + 1, the smaller the variations in frequency at each generation (Poiley 1960).

The degree of genetic heterogeneity in outbred colonies depends greatly on their history. It can be very low, for example, as a consequence of genetic drift (or the bottleneck effect), when the pool of breeders has been accidentally or intentionally reduced to a few individuals (this is common when a new breeding facility is created and a small group of breeders is imported). In contrast, genetic heterogeneity can be much higher when the stock has been recently outcrossed. Some commercial breeders probably monitor the polymorphisms segregating in their stocks with DNA markers, but the methodology they use and the results they get are not always made public. Being genetically heterogeneous, outbred and randombred stocks have a greater fertility index than inbred strains and, accordingly, they are sold at a much lower price per unit.

Because outbred colonies are heterogeneous populations, like human populations, they are often considered as being the most appropriate category of laboratory animals to use in toxicology, and pharmacology research. However, several geneticists have disputed this point of view and it has even been considered that, in many studies, outbred mice were used inappropriately, wasting animals' lives and resources on suboptimal experiments (Chia et al. 2005; Festing 2010). In fact, any outbred stock can be replaced by a "synthetic" population obtained by intercrossing classical inbred strains. As we already said, crossing two inbred strains to produce an F1 progeny and then crossing two independent F1 generates a four-way polymorphic population. This population is *heterogenic*, in the sense that individuals are genetically different.

In addition, the population often carries a greater number of allelic forms, which is generally considered an advantage compared to a classical outbred population.

Recently, however, researchers have considered that outbred stocks might be useful to refine the identification of QTLs, because these heterogeneous stocks accumulate in their genome many recombination breakpoints over time that split their chromosomes into "fine-grained mosaics", facilitating the high-resolution mapping of complex traits (Mott et al. 2000; Flint et al. 2005; Yalcin et al. 2010).

Finally, random-bred stocks are of very limited interest to geneticists. These stocks are bred with no specific rules, paying almost no attention to the genetic diversity in the population. Since they are in general of relatively small size, they drift rapidly towards a moderately inbred but still undefined population.

References

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, Ferris MT, Frelinger JA, Heise M, Frieman MB, Gralinski LE, Bell TA, Didion JD, Hua K, Nehrenberg DL, Powell CL, Steigerwalt J, Xie Y, Kelada SN, Collins FS, Yang IV, Schwartz DA, Branstetter LA, Chesler EJ, Miller DR, Spence J, Liu EY, McMillan L, Sarkar A, Wang J, Wang W, Zhang Q, Broman KW, Korstanje R, Durrant C, Mott R, Iraqi FA, Pomp D, Threadgill D, de Villena FP, Churchill GA (2011) Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res 21:1213–1222

Bailey DW (1971) Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. Transplantation 11:325–327

Banbury (1997) Mutant mice and neuroscience: recommendations concerning genetic background: banbury conference on genetic background in mice. Neuron 19:755–759

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM (2000) Genealogies of mouse inbred strains. Nat Genet 24:23–25

Benavides FJ (1999) Genetic contamination of an SJL/J mouse colony: rapid detection by PCR-based microsatellite analysis. Contemp Top Lab Anim Sci 38:54–55

Berning AK, Eicher EM, Paul WE, Scher I (1980) Mapping of the X-linked immune deficiency mutation (xid) of CBA/N mice. J Immunol 124:1875–1877

Bishop CE, Boursot P, Baron B, Bonhomme F, Hatat D (1985) Most classical Mus musculus domesticus laboratory mouse strains carry a Mus musculus musculus Y chromosome. Nature 315:70–72

Bonhomme F (1986) Evolutionary relationships in the genus *Mus*. Curr Top Microbiol Immunol 127:19–34

Bonhomme F, Guénet JL (1996) The laboratory mouse and its wild relatives. In: Lyon M, Rastan S, Brown DM (ed) Genetic variants and strains of the laboratory mouse. Oxford University Press, New York. pp 1577–1596

Bonhomme F, Guénet JL, Dod B, Moriwaki K, Bulfield G (1987) The polyphyletic origin of laboratory inbred mice and their rate of evolution. Biol J Linn Soc 30:51–58

Bryda EC, Riley LK (2008) Multiplex microsatellite marker panels for genetic monitoring of common rat strains. J Am Assoc Lab Anim Sci 47:37–41

Bulfield G, Siller WG, Wight PA, Moore KJ (1984) X chromosome-linked muscular dystrophy (mdx) in the mouse. Proc Natl Acad Sci USA 81:1189–1192

Burgio G, Baylac M, Heyer E, Montagutelli X (2009) Genetic analysis of skull shape variation and morphological integration in the mouse using interspecific recombinant congenic strains between C57BL/6 and mice of the mus spretus species. Evolution 63:2668–2686

Burgio G, Szatanik M, Guénet JL, Arnau MR, Panthier JJ, Montagutelli X (2007) Interspecific recombinant congenic strains between C57BL/6 and mice of the Mus spretus species: a powerful tool to dissect genetic control of complex traits. Genetics 177:2321–2333

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet 10:783–796

Chen S, Kadomatsu K, Kondo M, Toyama Y, Toshimori K, Ueno S, Miyake Y, Muramatsu T (2004) Effects of flanking genes on the phenotypes of mice deficient in basigin/CD147. Biochem Biophys Res Commun 324:147–153

Chesler EJ, Miller DR, Branstetter LR, Galloway LD, Jackson BL, Philip VM, Voy BH, Culiat CT, Threadgill DW, Williams RW, Churchill GA, Johnson DK, Manly KF (2008) The collaborative cross at Oak ridge national laboratory: developing a powerful resource for systems genetics. Mamm Genome 19:382–389

Chia R, Achilli F, Festing MF, Fisher EMC (2005) The origins and uses of mouse outbred stocks. Nat Genet 37:1181–1186

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques AC, Graves T, Zhou S, Teague B, Potamousis K, Churas C, Place M, Herschleb J, Runnheim R, Forrest D, Amos-Landgraf J, Schwartz DC, Cheng Z, Lindblad-Toh K, Eichler EE, Ponting CP (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7:e1000112

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Paloti L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, Tarantino LM, Threadgill D, Toth LA, Valdar W, de Villena FP, Warden C, Whatley S, Williams RW, Wiltshire T, Yi N, Zhang D, Zhang M, Zou F, The Complex Trait Consortium (2004) The collaborative cross: a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137

Coleman DL, Hummel KP (1973) The influence of genetic background on the expression of the obese (Ob) gene in the mouse. Diabetologia 9:287–293

Davisson MT (1996) Rules for nomenclature of inbred strains. In: Lyon MF, Rastan S, Brown SDM (eds) Genetic variants and strains of the laboratory mouse. Oxford University Press, Oxford, pp 1532–1536

Dejager L, Libert C, Montagutelli X (2009) Thirty years of Mus spretus: a promising future. Trends Genet 25:234–241

Demant P (2003) Cancer susceptibility in the mouse: genetics, biology and implications for human cancer. Nat Rev Genet 4:721–734

Demant P, Hart AA (1986) Recombinant congenic strains–a new tool for analyzing genetic traits determined by more than one gene. Immunogenetics 24:416–422

Doetschman T (2009) Influence of genetic background on genetically engineered mouse phenotypes. Methods Mol Biol 530:423–433

Ferris SD, Sage RD, Wilson AC (1982) Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. Nature 295:163–165

Festing MF (1979) Inbred strains in biomedical research. Macmillan, London

Festing MF (2010) Inbred strains should replace outbred stocks in toxicology, safety testing, and drug development. Toxicol Pathol 38:681–690

Flint J, Valdar W, Shifman S, Mott R (2005) Strategies for mapping and cloning quantitative trait genes in rodents. Nat Rev Genet 4:271–286

Forejt J (1996) Hybrid sterility in the mouse. Trends Genet 12:412–417

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, Pethiyagoda CL, Stuve LL, Johnson FM, Daly MJ, Wade CM, Cox DR (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448:1050–1053

Freeman D, Lesche R, Kertesz N, Wang S, Li G, Gao J, Groszer M, Martinez-Diaz H, Rozengurt N, Thomas G, Liu X, Wu H (2006) Genetic background controls tumor development in PTEN-deficient mice. Cancer Res 66:6492–6496

Glenister PH, Thornton CE (2000) Cryoconservation–archiving for the future. Mamm Genome 11:565–571

Goios A, Pereira L, Bogue M, Macaulay V, Amorim A (2007) mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res 17:293–298

Gregorova S, Divina P, Storchova R, Trachtulec Z, Fotopulosova V, Svenson KL, Donahue LR, Paigen B, Forejt J (2008) Mouse consomic strains: exploiting genetic divergence between Mus m. musculus and Mus m. domesticus subspecies. Genome Res 18:509–515

Grüneberg H (1952) The genetics of the mouse, 2nd edn. Martinus Nijhoff, The Hague

Guénet JL, Bonhomme F (2003) Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet 19:24–31

Hartl DL (2001) Genetic management of outbred laboratory rodent populations. Charles River Genetic Literature

Hummel KP, Coleman DL, Lane PW (1972) The influence of genetic background on expression of mutations at the diabetes locus in the mouse. I. C57BL-KsJ and C57BL-6J strains. Biochem Genet 7:1–13

Johnson LL (1981) At how many histocompatibility loci do congenic mouse strains differ? J Hered 72:27–31

Kenneth NS, Younger JM, Hughes ED, Marcotte D, Barker PA, Saunders TL, Duckett CS (2012) An inactivating caspase 11 passenger mutation originating from the 129 murine strain in mice targeted for c-IAP1. Biochem J 443:355–359

Kuperwasser C, Hurlbut GD, Kittrell FS, Dickinson ES, Laucirica R, Medina D, Naber SP, Jerry DJ (2000) Development of spontaneous mammary tumors in BALB/c p53 heterozygous mice. A model for Li-Fraumeni syndrome. Am J Pathol 157:2151–2159

Linder CC (2001) The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases. Lab Anim (NY) 30:34–39

Mao HZ, Roussos ET, Peterfy M (2006) Genetic analysis of the diabetes-prone C57BLKS/J mouse strain reveals genetic contribution from multiple strains. Biochim Biophys Acta 1762:440–446

Markel P, Shu P, Ebeling C, Carlson GA, Nagle DL, Smutko JS, Moore KJ (1997) Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. Nat Genet 17:280–284

Mashimo T, Voigt B, Tsurumi T, Naoi K, Nakanishi S, Yamasaki K, Kuramoto T, Serikawa T (2006) A set of highly informative rat simple sequence length polymorphism (SSLP) markers and genetically defined rat strains. BMC Genet 7:19

Mattapallil MJ, Wawrousek EF, Chan CC, Zhao H, Roychoudhury J, Ferguson TA, Caspi RR (2012) The Rd8 mutation of the Crb1 gene is present in vendor lines of C57BL/6 N mice and embryonic stem cells, and confounds ocular induced mutant phenotypes. Invest Ophthalmol Vis Sci 53:2921–2927

Mattson DL, Dwinell MR, Greene AS, Kwitek AE, Roman RJ, Jacob HJ, Cowley AW Jr (2008) Chromosome substitution reveals the genetic basis of Dahl salt-sensitive hypertension and renal disease. Am J Physiol Renal Physiol 295:837–842

Mekada K, Abe K, Murakami A, Nakamura S, Nakata H, Moriwaki K, Obata Y, Yoshiki A (2009) Genetic differences among C57BL/6 substrains. Exp Anim 58:141–149

Moran N, Bassani DM, Desvergne JP, Keiper S, Lowden PA, Vyle JS, Tucker JH (2006) Detection of a single DNA base-pair mismatch using an anthracene-tagged fluorescent probe. Chem Commun 48:5003–5005

Moriwaki K, Shiroishi T, Yonekowa H (1994) Genetics in wild mice: its application to biomedical research. Japan Scientific Societies Press

Morse HC III (1978) Origins of inbred mice. Academic Press, New York

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci USA 97:12649–12654

Myakishev MV, Khripin Y, Hu S, Hamer DH (2001) High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. Genome Res 11:163–169

Nadeau JH, Singer JB, Matin A, Lander ES (2000) Analysing complex genetic traits with chromosome substitution strains. Nat Genet 24:221–225

Nijman IJ, Kuipers S, Verheul M, Guryev V, Cuppen E (2008) A genome-wide SNP panel for mapping and association studies in the rat. BMC Genom 9:95

Ogonuki N, Inoue K, Hirose M, Miura I, Mochida K, Sato T, Mise N, Mekada K, Yoshiki A, Abe K, Kurihara H, Wakana S, Ogura A (2009) A high-speed congenic strategy using first-wave male germ cells. PLoS ONE 4:e4943

Paigen K, Eppig JT (2000) A mouse phenome project. Mamm Genome 11:715–717

Petkov PM, Cassell MA, Sargent EE, Donnelly CJ, Robinson P, Crew V, Asquith S, Haar RV, Wiles MV (2004a) Development of a SNP genotyping panel for genetic monitoring of the laboratory mouse. Genomics 83:902–911

Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, Sargent EE, Asquith S, Crew V, Johnson KA, Robinson P, Scott VE, Wiles MV (2004b) An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 14:1806–1811

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K (2005) Evidence of a large-scale functional organization of mammalian chromosomes. PLoS Genet 1(3):e33

Poiley SM (1960) A systematic method of breeder rotation for non-inbred laboratory animals colonies. Proc Anim Care Panel 10:159

Poltorak A, He X, Smirnova I, Liu MY, Van Huffel C, Du X, Birdwell D, Alejos E, Silva M, Galanos C, Freudenberg M, Ricciardi-Castagnoli P, Layton B, Beutler B (1998) Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. Science 282:2085–2088

Rader K (2004) Making mice: standardizing animals for American biomedical research, 1900–1955. Princeton University Press, New Jersey

Rasmussen AL, Okumura A, Ferris MT, Green R, Feldmann F, Kelly SM, Scott DP, Safronetz D, Haddock E, LaCasse R, Thomas MJ, Sova P, Carter VS, Weiss JM, Miller DR, Shaw GD, Korth MJ, Heise MT, Baric RS, Manuel de Villena FP, Feldmann H, Katze MG (2014) Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. Science. pii: 1259595. [Epub ahead of print]

Schlager G, Dickie MM (1967) Spontaneous mutations and mutation rates in the house mouse. Genetics 57:319–330

Schuster-Gossler K, Lee AW, Lerner CP, Parker HJ, Dyer VW, Scott VE, Gossler A, Conover JC (2001) Use of coisogenic host blastocysts for efficient establishment of germline chimeras with C57BL/6 J ES cell lines. Biotechniques 31:1022–1026

Simon MM, Greenaway S, White JK, Fuchs H, Gailus-Durner V, Wells S, Sorg T, Wong K, Bedu E, Cartwright EJ, Dacquin R, Djebali S, Estabel J, Graw J, Ingham NJ, Jackson IJ, Lengeling A, Mandillo S, Marvel J, Meziane H, Preitner F, Puk O, Roux M, Adams DJ, Atkins S, Ayadi A, Becker L, Blake A, Brooker D, Cater H, Champy MF, Combe R, Danecek P, di Fenza A, Gates H, Gerdin AK, Golini E, Hancock JM, Hans W, Hölter SM, Hough T, Jurdic P, Keane TM, Morgan H, Müller W, Neff F, Nicholson G, Pasche B, Roberson LA, Rozman J, Sanderson M, Santos L, Selloum M, Shannon C, Southwell A, Tocchini-Valentini GP, Vancollie VE, Westerberg H, Wurst W, Zi M, Yalcin B, Ramirez-Solis R, Steel KP, Mallon AM, de Angelis MH, Herault Y, Brown SD (2013) A comparative phenotypic and genomic analysis of C57BL/6 J and C57BL/6 N mouse strains. Genome Biol 14(7):R82

Snell GD (1948) Methods for the study of histocompatibility genes. J Genet 49:87–108

Specht CG, Schoepfer R (2001) Deletion of the alpha-synuclein locus in a subpopulation of C57BL/6 J inbred mice. BMC Neurosci 2:11

Stevens JC, Banks GT, Festing MF, Fisher EM (2007) Quiet mutations in inbred strains of mice. Trends Mol Med 13:512–519

Strong LC (1978) Inbred mice in science in origins of inbred mice. In: Morse III HC (ed) Academic Press—Adapted for the Web by: mouse genome informatics. The Jackson Laboratory, Bar Harbor, Maine USA

Threadgill DW, Churchill GA (2012) Ten years of the collaborative cross. Genetics 190:291–294

Threadgill DW, Miller DR, Churchill GA, de Villena FP (2011) The collaborative cross: a recombinant inbred mouse population for the systems genetic era. ILAR J 52:24–31

Threadgill DW, Dlugosz AA, Hansen LA, Tennenbaum T, Lichti U, Yee D, LaMantia C, Mourton T, Herrup K, Harris RC et al (1995) Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype. Science 269:230–234

Tucker PK, Phillips KS, Lundrigan B (1992) A mouse Y chromosome pseudogene is related to human ubiquitin activating enzyme E1. Mamm Genome 3:28–35

Wade CM, Kulbokas EJ 3rd, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ (2002) The mosaic structure of variation in the laboratory mouse genome. Nature 420:574–578

Wakeland E, Morel L, Achey K, Yui M, Longmate J (1997) Speed congenics: a classic technique in the fast lane (relatively speaking). Immunol Today 18:472–477

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Wolfer DP, Crusio WE, Lipp HP (2002) Knockout mice: simple solutions to the problems of genetic background and flanking genes. Trends Neurosci 25:336–340

Wotjak CT (2003) C57BLack/BOX? The importance of exact mouse strain nomenclature. Trends Genet 19:183–184

Yalcin B, Fullerton J, Miller S, Keays DA, Brady S, Bhomra A, Jefferson A, Volpi E, Copley RR, Flint J, Mott R (2004) Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. Proc Natl Acad Sci USA 101:9734–9739

Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, Osterås M, Whitley A, Yuan W, Gan X, Goodson M, Klenerman P, Satpathy A, Mathis D, Benoist C, Adams DJ, Mott R, Flint J (2010) Commercially available outbred mice for genome-wide association studies. PLoS Genet 6(9):e1001085

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellaker C, Goodstadt L, Nicod J, Bhomra A, Hernandez-Pliego P, Whitley H, Cleak J, Dutton R, Janowitz D, Mott R, Adams DJ, Flint J (2011) Sequence-based characterization of structural variation in the mouse genome. Nature 477:326–329

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH, Nachman MW, Pialek J, Tucker P, Boursot P, McMillan L, Churchill GA, de Villena FP (2011) Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43:648–655

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Migita S, Bonhomme F, Hjorth JP, Petras ML, Tagashira Y (1982) Origins of laboratory mice deduced from restriction patterns of mitochondrial DNA. Differentiation 22:222–226

Zou F, Gelfond JA, Airey DC, Lu L, Manly KF, Williams RW, Threadgill DW (2005) Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. Genetics 170:1299–1311

Zurita E, Chagoyen M, Cantero M, Alonso R, Gonzalez-Neira A, Lopez-Jimenez A, Lopez-Moreno JA, Landel CP, Benitez J, Pazos F, Montoliu L (2011) Genetic polymorphisms among C57BL/6 mouse inbred strains. Transgenic Res 20:481–489

Chapter 10
Quantitative Traits and Quantitative Genetics

10.1 Introduction

In contrast with qualitative or dichotomous traits, quantitative traits are measured using quantitative or semi-quantitative variables and their inheritance is controlled by multiple genes acting independently or in association. Quantitative traits are also influenced to varying degrees by the environment and this explains why they are often designated complex traits, with the adjective "complex" referring more to the determinism of the phenotypic expression than to the trait itself.

Plasma cholesterol level, blood glucose level, daily water intake, body weight when adult, susceptibility to certain forms of cancer or to a particular infectious agent, etc. are all examples of quantitative traits because, while marked differences often exist between mice of different inbred strains, denoting an obvious genetic control for the traits in question, individual measures may also vary between animals of the same strain, although they are all genetically identical, indicating non-genetic sources of variation.

Because of these differences from qualitative traits, it has been suggested in the past that quantitative traits might obey non-Mendelian patterns of inheritance. Nowadays, it is established that the inheritance of quantitative traits is based on the same Mendelian principles as qualitative traits, but their inheritance cannot be analyzed with the same methodologies.

Understanding the mechanisms of inheritance of quantitative traits and ultimately identifying the genes that influence such traits is one of the major challenges geneticists must address nowadays because many human disorders with high prevalence (obesity, hypertension, cancers, etc.) as well as susceptibility to many common diseases in humans and animals are inherited as complex traits. Similarly, the selection of the best breeders in domestic animals is essentially based on a judicious exploitation of quantitative traits (Mackay 2009; Mackay et al. 2009). The genetics of complex traits is the topic of the present chapter, to which we will add a brief comment about the genes that have a modifying effect on the phenotypic expression of Mendelian traits.

This part of mammalian genetics makes use of various models for the analysis of the experimental data, often requiring a solid background in mathematics and

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6_10

statistics. In this chapter, we have deliberately chosen to minimize the mathematical developments. However, references to relevant textbooks or publications will be provided for those readers willing to gain a wider competence in the matter.

10.2 Mean and Variance: Two Essential Parameters for the Characterization of a Population

The main characteristic of quantitative traits is that, even when all known parameters influencing the trait and its measurement are perfectly controlled, trait values are still subject to inevitable and incontrollable fluctuations. This applies to repeated measurements of the same trait performed on the same individual, and to measurements performed on a group of individuals that share the same genetic makeup and environmental exposure from the moment of conception to the moment of analysis.

If one wishes to assess very robustly the phenotype of a particular individual, one must repeat the same measurement several times under exactly the same experimental conditions and use the mean of all values as the best estimate. For example, blood pressure can be measured on mice using an inflatable tail cuff but, although the reliability of this technique has dramatically improved, it is recommended to perform multiple measurements under the same conditions (same hour, same operator, same apparatus, etc.). The mean of all collected values will be used as the most accurate phenotypic assessment. The variance or its derivatives (standard deviation, SD; standard error of the mean, SEM) will provide a useful estimate of the repeatability of the measurement. These fluctuations generally remain in a quite narrow range and represent the individual variability reflecting some transient, within-animal changes not necessarily related to its metabolism.

The same issue arises at the population level. If one wishes to establish the blood pressure of male mice of the C57BL/6J inbred strain at 12 weeks of age under standard diet, it is absolutely necessary to make this measurement on a group of mice that have been bred under exactly the same conditions and environment. Values measured on individual mice of the group will be slightly different from one another and none of them can be taken as the "true estimate": this is referred to as inter-individual variations. The best estimate is again the mean of all values. In this case, it is usually not necessary to repeat the measurement several times on the same individuals, since the main source of variation will be between individuals. Here again, the variance is an important parameter for evaluating the fluctuations of the trait within a group of genetically and environmentally homogeneous individuals (Table 10.1 and Fig. 10.1).

These two types of variations are part of the concept called *residual variance*. Residual variance represents the part of the total variance of the individual values of a parameter, measured in a group of individuals, which cannot be explained either by differences in the genetic factors, by variations in the environmental factors, or by variations in the measurement methodology.

Table 10.1 The total plasma cholesterol levels of inbred mice

Strains	Cholesterol levels	SD
MOLF/EiJ	520	± 41.4
CAST/EiJ	429	± 62.4
NZB/BINJ	358	± 61.0
C3H/HeJ	301	± 18.9
SWR/J	220	± 14.4
129S1/SvlmJ	218	± 14.2
BALB/cJ	182	± 15.4
SJL/J	182	± 29.1
C57BL/6J	173	± 12.7
AKR/J	152	± 24.9
DBA/2J	119	± 12.7
CBA/J	101	± 09.1

The data on this table represent the mean plasma cholesterol levels (in mg/dL ± SD) in female mice of 12 commonly used inbred strains (aged 15–19 weeks). The inter-strain variations for this parameter are relatively large while the intra-strain variations are (in most cases) relatively limited. This indicates that this trait is clearly under genetic control. The data are from the Mouse Phenome database http://phenome.jax.org/db/qp?rtn=views/measplot&brieflook=2920&projhint=Paigen1

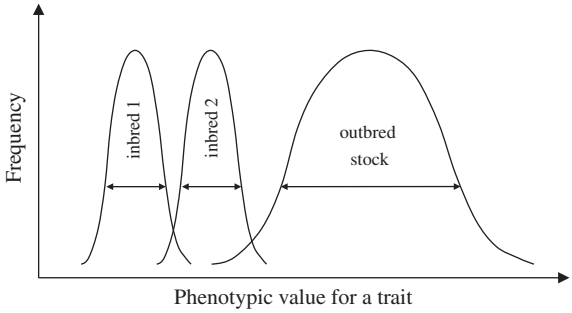


Fig. 10.1 *Distribution of a quantitative trait in inbred and non-inbred populations*. The phenotypic variance of a quantitative trait measured in a population of individuals is generally higher in outbred populations than in inbred strains due to genetic heterogeneity. As a consequence, inbred strains offer a greater power of resolution for detecting a difference in mean phenotypic value induced, for example, by a drug, a mutation or a transgene. The values observed in two different inbred strains sometimes partially overlap, illustrating the lack of absolute correlation between genotype and phenotype when dealing with quantitative characters

As a consequence of inter-individual fluctuations, it is frequently observed that two genetically different populations which show a different mean value for a trait exhibit a partially overlapping distribution of the phenotypic values. This

illustrates that not only is it impossible to assign precisely a trait value to a genotype, but it is also impossible to infer the genotype of an individual from its phenotype.

10.3 Why Study the Genetics of Complex Traits in Laboratory Mice?

Many of the quantitative traits studied in laboratory animals are directly related to human physiology and pathology (for example, hypertension, diabetes, obesity, etc.) and with the recent advances in human genetics and genomics, one could consider that these traits would be better studied in human populations, where the results can be directly exploited, rather than being explored in a model organism whose biology may differ from human biology. However, this approach is challenged by the difficulties inherent in the identification of genes controlling quantitative traits, especially if we remember that each trait is under the control of an unknown number of genetic factors whose effects are variable in intensity. In most instances, none of these genetic factors is sufficient, in itself, to induce the observed phenotype but each of them contributes, to some extent, to its expression. In addition, these multiple genetic factors are often involved in complex epistatic interactions making it difficult to tease apart their individual effects. Finally, and most importantly, environmental factors can modulate the biological effects of genetic factors, making their analysis even more complex.

Laboratory rodents, namely mice and rats, offer very potent means of analyzing the genetic control of complex traits in highly standardized and controlled conditions. Crossing inbred strains with established phenotypic differences offers the possibility of investigating gene–phenotype associations. Moreover, the existence of populations that are highly standardized from the genetic point of view, such as recombinant inbred strains, recombinant congenic strains, congenic strains or strains from the Collaborative Cross (described in Chap. 9), provide exceptional tools for gene detection and the evaluation of allelic effects. In addition, the genomic sequence is available for several strains of the mouse species, and a wide range of strategies is available to induce genetic alterations and study their effect, as described in the previous chapters. For all these reasons, the mouse offers unparalleled opportunities for exploring the genetics of quantitative traits of biomedical interest.

10.4 The Genetic Determinism of Quantitative Traits

Based on their determinism one can consider that quantitative traits are of two categories. The first category, the simplest, is when individual alleles that participate in the definition of a phenotype act independently by merely adding up the primary effect of each of them with no other form of interaction. This situation, which is rather rare, corresponds to what geneticists call the *additive model*. Figure 10.2

Individual effect of genotypes

	A/A	*A/B*	*B/B*
locus *1*	-5	0	+5
locus *2*	-8	+3	+7
locus *3*	-4	-2	+6

Mean phenotypic values for some genotypes

	Genotype	Value
1	$1^{A/A}\ 2^{A/B}\ 3^{A/A}$	-6
2	$1^{B/B}\ 2^{A/B}\ 3^{A/B}$	+6
3	$1^{A/A}\ 2^{B/B}\ 3^{A/A}$	-2
4	$1^{A/B}\ 2^{A/A}\ 3^{B/B}$	-2

Fig. 10.2 *Relationship between genotype and phenotype*. The figure represents a case of multigenic inheritance where the loci have only additive effects. In this example, a quantitative trait is controlled by three independent polymorphic loci (*1*, *2*, and *3*). The *left* panel indicates the effects of the genotypes at the three loci (*A/A, A/B* or *B/B*) on the average value in the population. The *right* panel shows the average phenotypic values associated with some (actually four) genotypic combinations, calculated as the (algebraic) sum of the effects of genotypes at each locus. One can see that genotypes *3* (*A/A; B/B; A/A*) and *4* (*A/B; A/A; B/B*), although different, are associated with the same average phenotypic value (-2). This example illustrates that one cannot infer the genotype of an individual from its phenotype as a quantitative trait

provides an example of such an additive effect, where a phenotype result from all possible combinations (9) of two alleles (*A* or *B*) at three loci (*1*, *2* or *3*). This example also illustrates the important notion that individuals with a different genetic make-up may nonetheless exhibit the same phenotype (ex: 3 and 4 in Fig. 10.2-right box). In this simple situation, the identification of genetic factors depends on the strength of gene effects and on the size of the population analyzed. In most cases of quantitative inheritance, the additive model does not explain all experimental observations and one must then make the assumption that *epistatic interactions* operate among the different genes with the effect of the different alleles at a given locus depending on the genotype at one or several other loci: a complex situation indeed!

10.5 The Concept of Quantitative Trait Locus (QTL)

The genetic determinants that are responsible for quantitative traits are in general numerous and, for this reason, they have been designated polygenes in the past. Nowadays, they are known as *quantitative trait loci* (QTLs). A QTL is defined as a locus or haplotype whose different alleles are associated with different average phenotypic values. For example, if individuals homozygous for the *a* allele at locus *X* (X^a/X^a) are on average significantly heavier than those which are homozygous for the *b* allele (X^b/X^b) (in the absence of any other difference between the two groups, such as sex, age, food, genetic background, etc.), we can conclude that there is, at locus *X* or in its vicinity, a gene that controls body weight. Locus *X* is called a QTL, a locus controlling a quantitative trait.

Note that this effect can be assessed only on groups of individuals since, once again, no conclusion can be drawn from single individuals. The difference between the body weight of the animals differing in their genotype at locus *X* must be statistically

significant, which may require using large groups of animals if the effect of the QTL (the body weight difference) is small. Using fewer animals may not reveal this difference and the QTL may be missed. Therefore, the capacity to detect QTLs is directly related to the experimental design, in particular to the number of animals analyzed, as well as to the strength of the QTLs segregating in the population.

Most quantitative traits are determined by several QTLs with a wide range of effect size, and these QTLs together control part of the phenotypic variation in a population. As previously mentioned, environmental parameters also contribute to this variation, as well as other sources, namely the interactions between the genotype of an individual and its environment (often designated G×E, reflecting the fact that genes' effects can vary in different environments). Finally, uncontrolled errors in measuring the phenotype of interest can also occur.

When deciphering the genetic control of a trait, it is important to quantify the contribution of genetic factors to the phenotypic variations. *Heritability* measures this contribution and is defined as the ratio of genotypic variance to phenotypic variance in the population that was analyzed. It has been refined into two more precise estimates. *Broad-sense heritability* takes into account the variance due to all types of genetic effects: additive effects, dominance effects, and epistatic interactions. *Narrow-sense heritability* considers only additive effects. Heritability (in both senses) is therefore a variable between 0 and 1. Higher values correspond to traits that are under stronger dependence of genetic factors. For example, the heritability of a fully penetrant Mendelian mutation is 1. Quantitative traits of medical significance have very variable heritability, and can be as low as 0.2. It is important to note that the heritability value is not an intrinsic characteristic of a trait, but depends on the population from which it was estimated, since it is conditioned by the number and nature of genetic variants segregating in this population.

10.6 Positioning QTLs on the Genetic Map

The genetic mapping of genes controlling a quantitative trait is based on the identification of differences in the average phenotype between groups of individuals depending on their genotype at a particular genomic location. Although this may resemble the procedure used for the mapping of qualitative traits, there are however important differences that result from the poor genotype–phenotype correlation at the individual level.

A first major difference is that, since the genetic alteration causing a Mendelian or qualitative trait generally involves only one locus, the genotyping of progeny can be interrupted when significant evidence of linkage has been detected between the locus of the mutant allele and one or a few flanking markers. For the localization of quantitative traits the situation is radically different because, in general, one does not know the number of QTLs involved in the determinism of the phenotype and for this reason the genotyping of a progeny must be carried out until the entire genome is covered with many evenly spaced markers.

Second, while strategies have been developed to identify the chromosome carrying a mutation using less than a few dozen animals, QTL mapping always requires the analysis of a large cohort, of at least one or two hundred animals, usually resulting in relatively imprecise location with large confidence intervals.

While the fine mapping of a Mendelian mutation can be achieved through the identification of individuals whose chromosomes have recombined between a genetic marker and the mutant locus, refining the location of a QTL requires the analysis of a group of animals carrying the same haplotype in order to estimate their average phenotype.

Finally, although data analysis is straightforward for qualitative traits segregating in crosses between two inbred strains, QTL detection requires sophisticated statistical models to account for gene interactions, especially when they are complex. For these reasons, QTL mapping, and even more so QTL characterization, is a much more difficult and lengthy endeavor than finding the gene responsible for a Mendelian trait.

10.6.1 Using Two-Generation Crosses for the Detection and Positioning of QTLs

The crosses that are most frequently used for the genetic localization of mouse QTLs are backcrosses (BC) or F2 intercrosses (F2) bred from parental inbred strains. Most of the time, these crosses involve strains where large phenotypic differences exist for the trait being measured. This would be the case, for example, of a cross between a hypertensive and a normotensive strain or between any two strains with marked differences in daily food intake. This situation applies to any phenotype that can be measured in individual animals using a quantitative variable. In some cases, several traits are measured on each animal, as a way of better describing the status of the individual for a given condition. Each trait can then be submitted to genetic analysis independently. Alternatively, several traits can be combined into a composite variable derived from mathematical combination of the original measurements to challenge the hypothesis that genetic (epistatic) interactions possibly occur.

A special situation applies to the genetic predisposition to develop a certain disease. Recording only the death or the absence of the disease in every animal is a binary trait that is poorly informative. A better quantitative measurement would be the age of the onset of the disease and a much more refined evaluation of the susceptibility would be to measure phenotypes at the cell or the organ levels that reflect pathophysiological processes characteristic of the disease.

In the mouse, unfortunately, not many strains spontaneously develop a disease faithfully modeling a homologous human condition. When they exist, these strains have been crossed with a wide variety of normal (resistant or healthy) strains for the purpose of QTL mapping. An example is the NOD (non-obese diabetic) strain of mice, which spontaneously develops type I diabetes mellitus. This strain has been frequently used to study the genetic determinism and pathology of diabetes in crosses with a variety of diabetes-resistant strains (e.g., the non-obese normal inbred strain or NON).

The choice of the parental strains for making a particular cross has a major influence on the number and position of the QTLs that will be identified. Whatever the situation, the greater the phenotypic differences, the greater the chance of detecting QTLs involved in the determinism of the trait being studied.

Choosing the most appropriate type of cross (BC or F2) is another important decision that is often guided by the phenotype of the F1s. One has also to take into account the interactions (dominant, recessive, additive or epistatic) of the different alleles. In some circumstances it may be wiser to analyze an F2 rather than a backcross because, in this case, all sorts of genotypes (*a/a, a/b, b/b*) appear in the progeny. In a backcross progeny, on the other hand, only two classes of genotypes occur, *a/a* and *a/b,* and this may hamper the detection of a QTL in which the *b* allele would be recessive.

Before making the cross, it is also very important to establish the mean value and the variance of the phenotype for the two parental strains and their F1. Since all the mice within each parental strain or within their F1s are genetically homogeneous, the observed variances should be of the same order of magnitude in the three groups since phenotypic variations originate from non-genetic factors. In all cases, the knowledge of the average values of the different inbred strains and their F1 progeny is important for deciding the best cross to make. When the average value of the F1 is close to the average value of one of the parental strain, it is recommended to make a backcross by crossing the F1 with the other parental strain. When the average value of the F1 is intermediate, deciding on the F2 is a sound choice (Fig. 10.3).

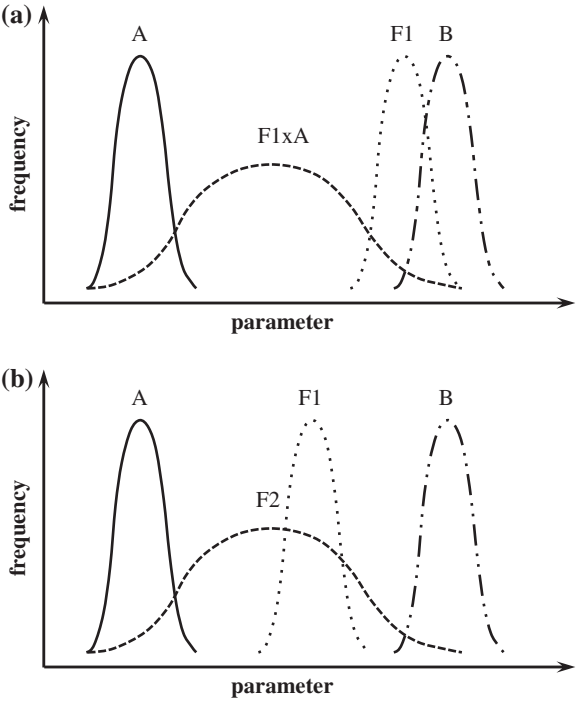


Fig. 10.3 *Distribution of a quantitative trait in a cross between two inbred strains*. The phenotypic variance in the F1 is of the same order of magnitude as that observed in the parental lines, due to the lack of genetic variability in these groups. **a** If the average phenotypic value of the F1 is close to one of the two parental lines (line B in this case) then it would be more informative to breed a backcross progeny by crossing with strain A (F1 × A). **b** If the average phenotypic value of the F1 is intermediate between those of the two parental strains, choosing to breed an F2 is a better option

10.6.2 Point-by-Point Analysis of the Progeny

The F2 or BC progeny bred for the genetic localization of QTLs must be carefully phenotyped using the same protocol as for the parental strains. Any increase in the phenotypic variance in the BC or F2 population—compared with that of the parental strains and F1—results by definition from an increase in the genetic variability of the population in question and reflects the action of the genetic factors segregating in the cross on the phenotypic variance.

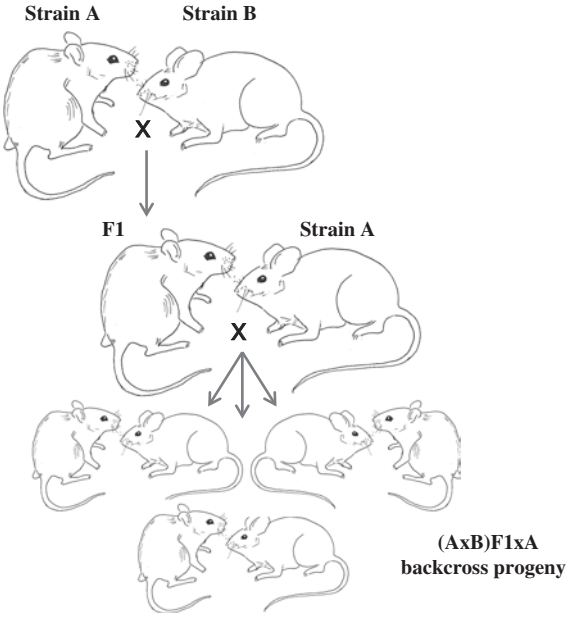
Phenotyping must also be performed in a very standardized manner because, if genotyping errors can be detected when analyzing data and easily corrected by retyping, phenotypic values can in general be assessed only once on every animal, and no longer after its death, should it occur. Accurate phenotyping counts at least as much as accurate genotyping in the mapping of QTLs.

The genotypes of the animals are established by typing genetic markers evenly distributed over the genetic map. Nowadays, these markers are microsatellites or SNPs selected in order to achieve an average spacing of 10–15 cM. Before any QTL mapping analysis is performed, it is highly recommended to check that the observed genotypes are consistent with the known position of the markers on the chromosome map. Some computer programs offer features for detecting genotyping errors.

The first level of analysis consists of seeking an association between each genotyped marker and the phenotype. The aim is to identify markers for which individuals carrying different genotypes (*a/a* or *a/b* in a backcross; *a/a*, *a/b* or *b/b* in an F2) show different average phenotypes. For each marker, the offspring are sorted according to their genotype and the mean phenotypic values of the different genotypic classes are compared using Student's *t*-test or analysis of variance (ANOVA) (if the phenotypic values follow a normal distribution, either as raw values or after appropriate transformation) or a non-parametric test. A significant difference suggests the existence of a QTL in the vicinity of the marker (Fig.10.4). By repeating this analysis for all markers genotyped, one can identify all chromosomal regions that are playing a role in the genetic control of the trait. In a given chromosomal region, the QTL is most likely located close to the marker with the strongest association (based on the *p*-value). To refine the likely position of the QTL, additional markers can be genotyped in the region of interest, but it is not helpful to perform mapping with a high density of markers in a backcross or an F2 population (5 cM spacing is sufficient).

10.6.3 The Concept of LOD Score

To define the statistical significance of a QTL, geneticists use the LOD score (*logarithm of the odds*). The LOD score is a statistic that compares the likelihoods of two alternative hypotheses referring to the phenotypic difference observed between two classes of genotypes at a particular marker. The first hypothesis is that the observed difference is indeed due to the presence of a QTL in the vicinity



Markers	Genotype		Student's	p
	a/a	*a/b*	*t test*	
Locus *X*	127 ± 41 (N = 45)	132 ± 38 (N = 53)	0.63	> 0.5
Locus *Y*	122 ± 29 (N = 51)	140 ± 27 (N = 47)	3.17	0.002

Fig. 10.4 *Pointwise statistical analysis in the case of a backcross between two strains*. The backross progeny was produced by crossing F1 with strain A. The phenotypic values in the backcross population were distributed according to a Gaussian (normal) distribution. Genotyping was then achieved by typing the backcross individuals for marker loci whose position is known and evenly distributed over all chromosomes. To test the effect of the genotype at a given locus, the average phenotypic values of homozygous (*a/a*) and heterozygous (*a/b*) offspring at this locus were compared using Student's t test. Here, we compare the average phenotypic values of individuals homozygous (*a/a*) and heterozygous (*a/b*) for two loci X and Y. In the case of locus X, there is no significant difference between the two groups. In contrast, mice homozygous *a/a* at the Y locus exhibit an average phenotypic value significantly lower (122 ± 29) than that of heterozygous individuals (*a/b*) (140 ± 27). We conclude that there is a QTL controlling the trait studied in the proximity of the Y locus

of the marker while the second hypothesis considers that the difference results only from random fluctuations. The LOD score computes the ratio between the likelihoods of these two hypotheses and expresses this ratio as a base-10 logarithm. The higher the LOD score, the more likely the presence of a QTL in the region in question. A LOD score value of 3 calculated for a marker indicates that the association between the phenotype and the genotype at this marker is 10^3 (1,000) times more likely to be due to the existence of a QTL close to this marker than to random fluctuations.

10.6.4 Threshold of Significance

Determining the actual level of statistical significance when performing QTL mapping is an issue. If 200 genetic markers have been genotyped, 200 statistical tests will be performed and there is a risk that some of them will lead to *p*-values below the standard 0.05 threshold just by chance. In fact, this level of significance means that the difference observed could happen in 1 out of 20 tests by chance, i.e. in the absence of any effect of the marker on the phenotype. With 200 markers tested, one would expect to get 10 markers associated by chance with the phenotype with a *p*-value of 0.05, in the absence of any true QTL. Therefore, a more stringent threshold must be adopted to avoid these false positives.

An abundant literature has addressed this issue. Appropriate significance levels depend on the type of cross, phenotype distribution, and marker density. Nowadays, it is generally accepted that the optimal strategy for estimating significance thresholds is phenotypic data permutation. Animal genotypes remain unchanged but phenotypic data are reshuffled between animals, to break all true causative genotype–phenotype associations. When permuted data are submitted to QTL analysis, all detected associations are false-positive. For each permutation, the highest LOD score observed is considered. By performing hundreds or thousands of such permutations, one can calculate the frequency at which LOD scores of 3, 4, 5, etc. were observed, all of which are false positives. One can also determine the LOD score value that has been observed in exactly 5 % of the permutations. This LOD score value is taken as the true 0.05 threshold. All recent QTL mapping programs incorporate data permutation.

10.7 Assessing the Strength of a QTL on the Trait Studied

Once a QTL has been identified close to a genetic marker, one can evaluate the strength of its effect on the trait by calculating the proportion of the phenotypic variance controlled by the QTL in the population studied. We will consider the case where the effects of the genotypes are not influenced by environmental factors (no gene × environment interactions).

The total phenotypic variance (V_T) in a F2 or backcross (BC) population is the sum of the phenotypic variance of genetic origin (V_G) and of the phenotypic variance due to individual and environmental factors (V_E). V_E can be estimated by the phenotypic variance measured in the parental lines or in the F1. V_G is therefore the difference between the phenotypic variance of the F2 or backcross population and the phenotypic variance of the F1 population.

When considering a particular marker, V_G can be decomposed into two fractions: V_Q, the genetic variance explained by the genotype at the marker, and V_{NQ}, the genetic variance explained by other genetic factors (Table 10.2). These two fractions are calculated from an ANOVA with the genotype at the marker as the main factor. The higher the V_Q/V_G ratio, the stronger the effect of the QTL.

Table 10.2 Origin of the phenotypic variance in different populations (the case of a backcross)

Population	Origin of Variance
Strain A	V_E
Strain B	V_E
F1	V_E
Backcross	$V_T = V_E + V_G$
Group of X^a/X^a offspring	$V_E + V_{NQ}$
Group of X^a/X^b offspring	$V_E + V_{NQ}$
Between X^a/X^a and X^a/X^b	V_Q

Animals are classified into two groups: X^a/X^a and X^a/X^b, according to their genotype at the marker locus X near which a QTL has been detected. V_E is the variance due to individual and environmental factors; V_T is the total variance in the backcross population; V_G is the variance of genetic origin in the backcross population; V_Q is the part of the genetic variance explained by the QTL; V_{NQ} is the fraction of the genetic variance due to other genetic factors (other QTLs). $V_G = V_Q + V_{NQ}$

10.8 Interval Mapping

The locus-wise analysis estimates the LOD score at each genotyped marker, i.e. the likelihood of the existence of a QTL at this position. These markers are usually separated by 10–15 cM and it is often useful to estimate the LOD score at intermediate positions. A simple approach would be to genotype additional markers to increase marker density. In fact, this is useful for the regions found to be associated with the phenotype, for refining the most likely position of a QTL, and this is the most accurate method. However, it is possible to interpolate genotypes between genotyped markers and compute LOD scores at intermediate position without genotyping additional markers. This method is called *interval mapping*.

Interval mapping consists of guessing the genotype of each animal at positions between two flanking markers, from the genotype of the animal at these markers and the recombination fractions between the position being assessed and the two flanking markers. Inferring the genotype at an intermediate position is straightforward when the two flanking markers are close and the animal has the same genotype at both markers. In this case, it is more than likely that the animal also carries the same genotype at all intermediate positions. In other cases, the algorithm considers all possible options, with their probability, to compute the LOD score. This results in maximizing the likelihood of existence of a QTL at this position. By performing this analysis at all positions between genotyped markers, one obtains a continuous LOD score curve for each chromosome. This curve is anchored at genotyped markers that provide reliable genotypes. LOD scores are less reliable at intermediate positions, and one should be very cautious if the flanking markers are separated by more than 20 cM. If a QTL is suspected in such a region, additional markers should be genotyped at intermediate positions.

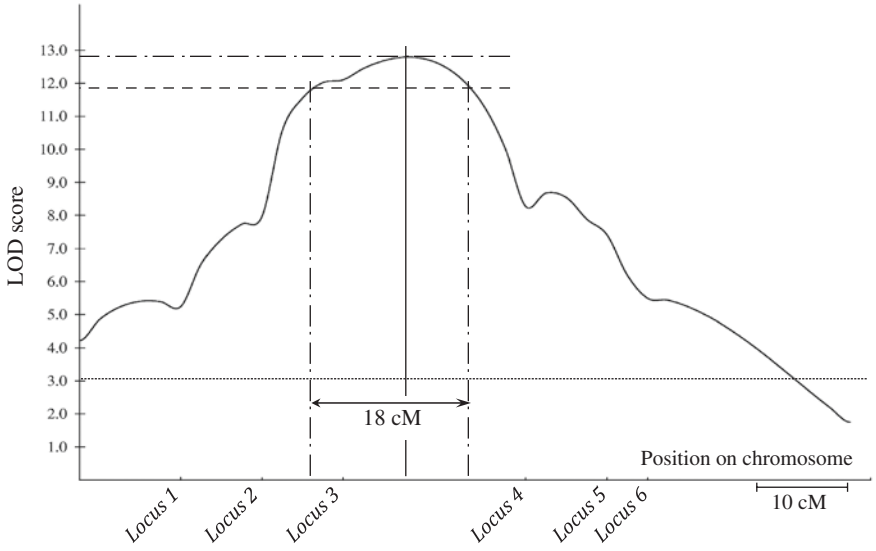


Fig. 10.5 *Determination of the confidence interval of a QTL using the LOD score curve established by interval mapping*. The X-axis represents the chromosome with the position of each analyzed marker (microsatellites or SNPs in general). The curve indicates the LOD score associated with the presence of a QTL at each position along the chromosome. The peak of the curve determines the position that represents the maximum likelihood for the presence of a QTL. The line corresponding to one $\log_{10}$ unit under the maximum LOD score (the second upper horizontal dotted line) is then drawn and the points of intersection of this line with the LOD score curve gives the confidence limits of the interval (18 cM in the case illustrated). In this case, it is recommended to genotype more markers in the QTL region to refine the LOD score curve and better define the confidence interval

The most likely position for a QTL is the one corresponding to the highest value for the LOD score (often called the peak of the curve) with a certain confidence level.

It is important to keep in mind that the existence of a QTL at a given position of the genetic map is associated with a certain probability of being right. However, in no way it is possible to conclude that a QTL exists with absolute certainty. As the LOD score falls below the significance threshold, the chances increase that the association is due to sampling fluctuations and not to the effect of a specific gene.

In the same way, the position of a QTL is not accurate. The precision of the positioning of a putative QTL along a chromosome is expressed as an interval that contains the QTL with a certain level of statistical confidence (for example 95 or 90 % confidence interval). Several methods exist to calculate the confidence interval (C.I.) associated to a QTL location. The simplest is based on the likelihood ratio test (Lander and Botstein 1989) and consists of moving sideward (left and right) of the estimated position to the locations corresponding to a decrease in the LOD score of either one or two units. The total width corresponding to one or two LOD drop-off can then be considered as the 96.8 or 99.8 % confidence interval, respectively. Another method uses Bayesian statistics and provides more relevant estimates (Fig. 10.5).

10.9 Searching for Multiple QTLs Simultaneously

The strategy outlined so far works under the assumption that each QTL is detectable independently from the others. However, there are frequent situations where the phenotypic effect of a QTL depends on the genotype of the animal at other genomic locations. In this case, scanning the genome one locus at a time misses these associations. Various statistical frameworks have been proposed to tackle this problem and fall into multiple QTL mapping approaches. One can use the genotype at a marker as a covariate in locus-wise analysis. For example, considering the genotype at a first QTL might help identifying others whose effect depends on the first QTL.

It is also possible to scan the genome for all pairs of genomic locations and test whether pairs of QTLs, acting either additively or in epistasis, can be detected. A number of models have been proposed and implemented in various statistical packages. This method can be time-consuming since many combinations of positions must be considered (especially when combined with interval mapping and data permutations). Moreover, because a huge number of tests are performed, one must use very stringent significance thresholds, which often precludes finding significant locus pairs. However, mapping QTLs controlling a trait cannot ignore the possibility of epistatic interactions.

10.10 Using Recombinant Inbred and Recombinant Congenic Strains

10.10.1 Recombinant Inbred Strains

One of the limitations of two-generation crosses (F2 or backcross) is that each progeny is unique, which does not allow replicating the assessment of the phenotype associated with a given genotype. Moreover, the density of recombination breakpoints among the experimental population, which ultimately determines the resolution of QTL mapping, is quite low. This has prompted many investigators to study the inheritance of complex traits by using recombinant inbred strains (RIS), which were described in Chap. 9. As discussed, each of these strains is inbred and accordingly all animals of a given strain are genetically alike. Given these conditions, one can establish the phenotype of each strain using an optimal number of individuals of each sex and thus obtain a mean phenotypic value and variance associated with a particular genetic make-up.

The strategy used for detecting QTLs when using RIS consists of the comparison of the average phenotypic value corresponding to one or the other parental allele at each locus whose genotype is established, then searching for possible genotype–phenotype associations. This comparison makes use of the ANOVA (analysis of variance) and compares the intra-strain variance (which is equivalent

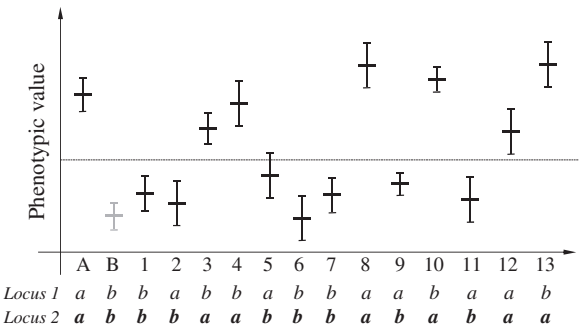


Fig. 10.6 *Using recombinant inbred strains for QTL location*. Parental strains A and B differ by a specific phenotypic trait. For this particular phenotypic trait a group of animals from each recombinant inbred strain has been phenotyped. The graph shows, for each strain, the mean value and standard deviation observed. For each marker analyzed, one looks (in general with the help of computer software) for a possible association between one parental allele and a high or low value for the phenotypic trait being studied. For locus 1, there is no obvious association between genotype and phenotype. However, for locus 2, strains that have a high average phenotypic value for the trait in question are all homozygous for the **a** allele, while those with low value are all homozygous for the **b** allele. Provided that the statistical test is significant, one can deduce the presence of a QTL for the trait studied in the vicinity of locus 2

to the residual variance since it concerns individuals which are all genetically identical), the inter-strain variance for those strains that have the same genotype at the marker locus (which is related to the genetic heterogeneity between the different lines), and the variance between the two groups of strains (which is controlled by the specific effect of the genotype at the marker locus) (Fig. 10.6).

10.10.2 Advantages and Disadvantages of RIS

The use of RIS offers certain advantages over two-generation crosses:

- Mice of RIS allow the very precise assessment of the phenotype associated with each genotype, since this assessment can be performed on a group of genetically identical animals. Replication improves the power of QTL detection and the precision of their positioning on the genetic map by reducing the phenotypic variance.
- Most RIS have already been genotyped for a great number of markers distributed throughout the whole genome; it is then only necessary to perform careful phenotyping for the trait under investigation.
- The genome sequences of most of the parental strains of existing RIS sets are now available. When these complete sequences are combined with haplotype reconstructions it is possible to inspect the full genome sequence for each RIS. This is a potentially very rich source of information.

- The genome of a RIS contains on average four times more recombination breakpoints than that of a backcross offspring, which provides finer resolution (reduced confidence interval) for QTL mapping (see Chap. 4).
- It is possible to compare results obtained in different laboratories on the same set of RIS, either in order to correlate two independent phenotypes or simply to assess the reproducibility of a study. Nothing like this can be done using the offspring of two-generation crosses that are all different.

However, RIS have some weaknesses:

- Sets of RIS exist only for a few pairs of parental inbred strains and the optimal set for the character under investigation may not be readily available. Note, however, that it is not necessary for the identification of QTLs that the two parental strains of an RIS set diverge for the phenotype of interest. In fact, the two parental strains may exhibit the same average phenotype though carrying different alleles at QTLs, while the strains of the recombinant inbred set may have inherited different associations of these alleles and then display a wide range of phenotypes, allowing for QTL identification and mapping. An example of this situation was published a few years ago (Grisel et al. 1997).
- When using RIS, the accurate location of a QTL depends on the number of strains analyzed, exactly as the accuracy of the location of a QTL depends on the number of offspring analyzed when using a two-generation cross. However, unlike in two-generation crosses, in which one can decide a priori the number of animals which will be analyzed and increase this number if desired, each set of RIS has a fixed number of strains that cannot be increased unless one decides to re-develop new RIS from the same parental strains and to genotype them, which is a considerable amount of work requiring additional crosses and technical investment. However, one can study several sets of RIS for the same phenotype and compare the QTLs identified. When common QTLs are found in two sets, data from the two sets can be combined and the mapping resolution is consequently increased.
- The analysis of the allelic interactions at a given locus (i.e. assessing the effects of recessivity/dominance between alleles) requires making additional crosses, whereas this is not the case in an F2 which features animals from all three classes of genotypes (*a/a, a/b,* and *b/b*) at each marker locus.

10.10.3 Recombinant Congenic Strains

Understanding the genetic control of complex traits require to isolate and analyze the individual effects of every QTL identified. When a character is controlled by several QTLs, RIS have a limitation that is a direct consequence of the equal contribution of both parental genomes in the genome of each RIS. Indeed, each RIS differs from its two inbred parental strains, on average, by half of the QTLs that are segregating. If we assume that a given trait is controlled by six independent QTLs, a given RIS will differ on average from each of its parental strain by three QTLs. In these circumstances it is difficult to study the individual effect of one particular QTL.

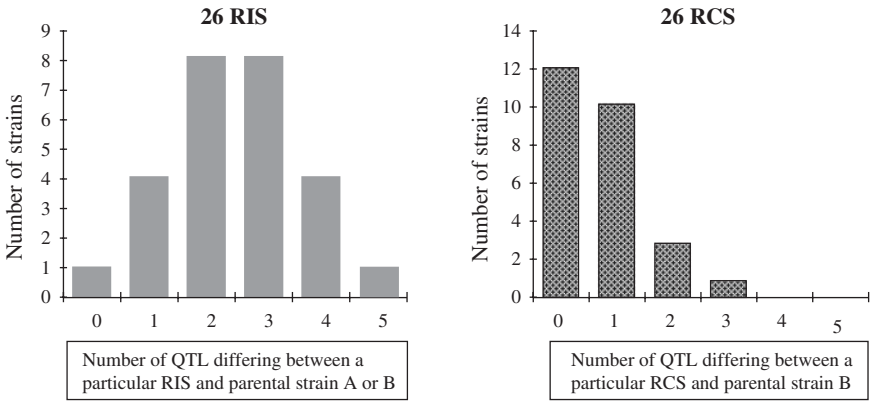


Fig. 10.7 *Advantage of recombinant congenic strains (RCS) for QTL analysis*. Comparison between a set of 26 RIS and a set of 26 RCS derived from the same parental strains A and B. If we consider that five QTLs control a characteristic making strain A different from strain B, around 16 (8 + 8) of the 26 RIS differ from the parental line A (or B) by two or three of these five QTLs and only four differ by a single one. On the other hand, 10 out of the 26 RCS differ from the background strain B by a single QTL. It is therefore easier to analyze the individual effects of each of the five QTLs by using RCS

It is mainly for this reason that other genetic populations, called *recombinant congenic strains* (RCS), were developed (Groot et al. 1992). These strains are also inbred, just like RIS, and they are derived from the offspring of a cross in which a donor inbred strain is previously backcrossed two or three times with another inbred line considered as recipient (see Chap. 9).

Each RCS differs from the background strain for one eighth (12.5 %) of the genome of the donor line. Using such RCS, it is then easier to find at least one strain that differs from the recipient strain by only one QTL, which allows assessment of its individual effect on the phenotype. Several sets of RCS have been developed for the analysis of complex traits and have been genotyped with hundreds of genetic markers. However, they have never reached the same popularity as RIS (Fig. 10.7).

10.11 Using Congenic Strains

When the phenotypic analysis of the progeny of a cross suggests the presence of a QTL in a particular chromosomal region, other experiments are required. First, it is necessary to confirm its existence since the presence of the QTL is not certain but only associated with a certain probability (assessed by the LOD score). Its position and the boundaries of the candidate interval must also be refined because the methods used for QTL detection result in ill-defined edges. Finally, and most importantly, the QTL should be isolated in a specific strain to assess its individual effects with no interference from the other QTLs possibly segregating in the same cross.

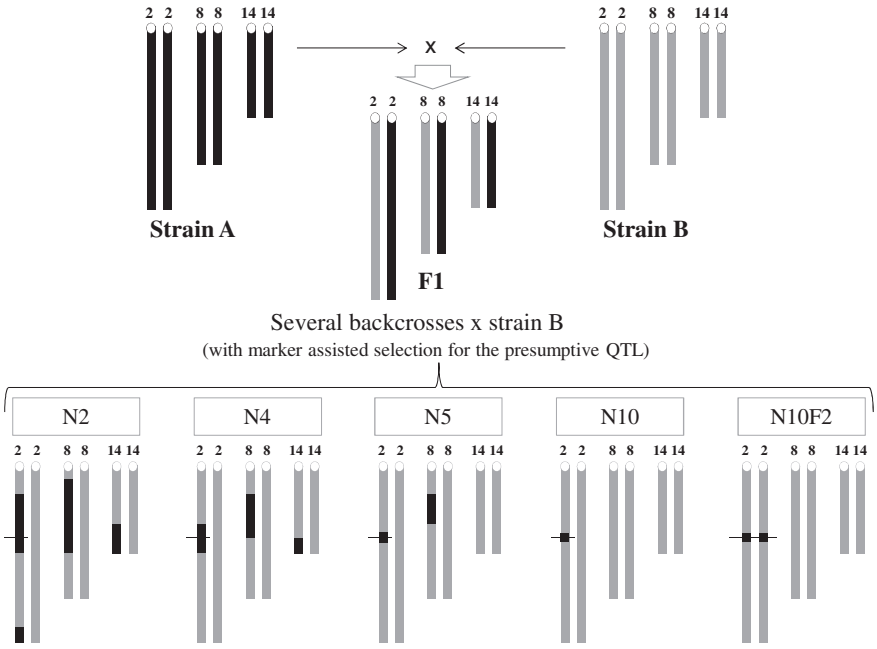


Fig. 10.8 *Congenic strains*. A congenic strain (described in Chap. 9) is different from the background strain by only a short chromosomal segment containing the presumptive QTL. This is achieved by selectively introgressing the chromosomal segment expected to contain the QTL from the parental strain A into parental strain B. At every generation, progeny are genotyped for a few markers flanking the confidence interval of the QTL, and animals heterozygous at all markers are kept for further breeding. Note that, if these two markers are distant by more than 15–20 cM, it is recommended to genotype additional markers within the interval to make sure that the animals selected for further breeding have retained the entire interval. The speed congenics strategy (see Chap. 9) can be used to accelerate the process of congenic strain development

To reach these three objectives, geneticists breed a strain congenic for this QTL. A congenic strain (as described in Chap. 9) is different from the background strain by only a chromosomal segment containing the presumptive QTL. This is achieved by introgressing selectively the chromosomal segment expected to contain the QTL from the parental strain B into parental strain A. At every generation, progeny are genotyped for a few markers flanking the confidence interval of the QTL, and animals heterozygous at all markers are kept for further breeding (Fig. 10.8). Note that, if these two markers are distant by more than 15–20 cM, it is recommended to genotype additional markers within the interval to make sure that the animals selected for further breeding have retained the entire interval. The *speed congenics* or *high-speed congenics* strategies can be used to accelerate the process of congenic strain development (see Chaps. 2 and 9).

To confirm a QTL, one can either develop a strain congenic of parental strain A for the B allele at the QTL (denoted A.B-QTL1) or the opposite (B.A-QTL1). In some cases, it may be useful to produce both.

If the parental strain A and its congenic partner A.B-QTL1 show a distinct phenotype, it is possible to conclude that the genomic region transferred in the congenic strain harbors one or more genetic factors controlling the trait. However, it should be noted that the absence of difference does not rule out the existence of a QTL in the region. Once isolated in the A background, the effect of the B allele may be too weak to change the phenotype of strain A. This is observed for example in the case where a trait is controlled by several QTLs with moderate individual effects. The analysis of congenic strains may fail to confirm these QTLs. In this case, it is recommended to intercross the congenic strains to produce strains carrying B alleles at two QTLs simultaneously.

One can then refine the location of the QTL by again crossing the A-QTL1^{-B} strain with strain A to break the original interval into a collection of smaller, partially overlapping, sub-intervals (Fig. 10.9). Comparing the phenotype of each of these sub-congenic strains with their genetic structure provides strong evidence to narrow down the location of the QTL. This process can be repeated to reduce as much as possible the size of the physical interval harboring the QTL.

Congenic strains are invaluable biological tools for investigating the nature, structure, and function of QTLs because they allow one to manipulate a single unit at a time. Among their many advantages, one is exceptional: working with congenic strains allows the study of the individual components of any trait. For example, diabetes or hypertension, which are two intensively studied complex traits

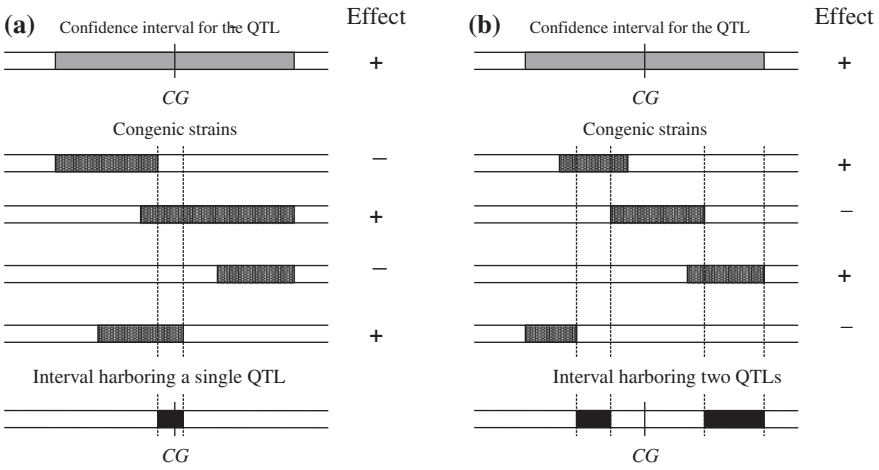


Fig. 10.9 *Using a series of overlapping sub-congenic strains to confirm and refine the location of a QTL.* **a** The picture represents a set of sub-congenic strains homozygous for a chromosome fragment (*grey rectangles*) of different size. By matching the chromosomal regions encompassed by the congenic segment with the phenotype of the different sub-congenic strains, with respect to the phenotype studied, it is possible to reduce the interval harboring a QTL. **b** The offspring of crosses set up between different sub-congenic strains (or RIS) are also useful for shortening the interval harboring a QTL. Sometimes this sort of experiment discloses the existence of two closely linked QTLs instead of only one. (*CG* = candidate gene)

in human, mice, and rats, result from the additive and interactive actions of an as yet undefined, although probably large, number of QTLs, having each a moderate effect (a sub-phenotype so to say). Isolating each of these QTLs in a congenic strain allows the study of their function and importance in the expression of the sub-phenotype even if the latter is modest.

Finally, congenic strains can be very useful for revealing the effect of weak QTLs otherwise masked by the strong effect of a major QTL. For example, susceptibility to Theiler's virus is strongly influenced by the major histocompatibility complex (*H2* locus). In a first cross between strains C57BL/10 and SJL/J, Brahic and Bureau (1998) identified this *H2* haplotype as a major factor. They made a second cross between SJL/J and the B10.S strain, which is congenic of C57BL/10 for the *H2* haplotype of the SJL strain. In this second cross, all progeny carry the same *H2* haplotype and the effect of other QTLs can be revealed more efficiently.

10.12 Using Other Strains or Stocks for the Mapping of QTLs

10.12.1 Consomic Strains (CS)

Consomic strains or chromosome substitution strains (CSS) are similar to congenic strains with the exception that the introgressed genomic segment is a complete chromosome (Singer et al. 2004). They are useful tools for the identification and analysis of QTLs but, unfortunately, only a few sets are available. When it is possible to use such a set, it allows rapid association of a phenotype with a particular chromosome for which the donor and background strains differ. If, for example, one observes that strain C57BL/6J differs from its consomic partner strain C57BL/6J-Chr $1^{A/J}$ (carrying chromosome 1 from A/J on an otherwise C57BL/6J background) for a specific trait, this means that chromosome 1 contains at least one QTL controlling the trait.

Further characterization of a QTL identified on a chromosome requires the production of a series of sub-consomic strains (analogous to sub-congenic strains) to refine the location of the QTL(s) and evaluate its (their) individual effect(s). One major limitation of consomic strains is that they miss QTLs located on different chromosomes that operate in epistatic interaction.

Specific chromosome substitution strains (PSCSS) are a variant of consomic strains where one specific chromosome of the recipient mouse strain C57BL/6J has been substituted by the homologous counterparts from several different inbred strains and not just one as in the case of consomics. These PSCSS form a special population that has an identical genetic background to the recipient strain C57BL/6 and differs only in the donor chromosomes (Xiao et al. 2010).

10.12.2 The Collaborative Cross (CC): A Novel, Powerful Tool for Studying the Genetics of Complex Traits

The Collaborative Cross (commonly abbreviated CC) has been described in Chap. 9, along with the other kinds of genetically standardized strains (Threadgill et al. 2002; Churchill et al. 2004; Threadgill and Churchill 2012), and is already considered a next-generation tool for the analysis of complex traits.

The CC consists of a large panel of mouse RIS descending from the random and unique association of an equal proportion of the genome of eight unrelated founder strains, comprising five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, and NZO/H1LtJ) and three wild-derived strains representing the three major *Mus musculus* subspecies (CAST/EiJ, PWK/PhJ, and WSB/EiJ). To date, over 400 strains have been developed and a few dozen are considered complete (homozygosity > 90 %).

The CC panel shares many advantages with the RIS and has additional advantages related to the greater number of genetic polymorphisms segregating among the different parental strains, to the much greater number of independent strains that will be available, and to the higher number of recombination breakpoints per strains. Computer simulations have indicated for example that 500 RIS of the Collaborative Cross would be adequate to map a single additive locus that accounts for only 5 % of the phenotypic variation to within 0.96 cM. Even if this mapping resolution will not be sufficient to identify single genes unambiguously, it represents nevertheless a considerable leap forward in the power of resolution.

10.12.3 Interspecific Recombinant Congenic Strains (IRCS)

IRCS are a variety of the RCS (mentioned above) with parental strains belonging to two different mouse species. They were developed from the parental strains C57BL/6 (the background strain) and an inbred strain derived from the *Mus spretus* species (SEG/Pas) as the donor strain (Burgio et al. 2007, 2012). These strains are equivalent to RCS discussed above with, however, some important differences. First, the introgressed component is of very remote origin and accordingly contributes to an important amount of polymorphism. Second, the genomic contribution of each parental strain is very unequal since each IRCS strain carries up to eight SEG/Pas chromosomal segments with an average size of 11.7 Mb, totalizing 1.37 % of the genome. Finally, when adding up the individual contributions of all 55 strains the SEG/Pas genome covers 39.7 % of the total genome. IRCSs are useful to unravel QTL with small effects and gene interactions.

10.12.4 Diversity Outbred (DO) Stock

The major limitation in QTL mapping is the resolution, and resolution itself depends on the density of recombination breakpoints in the individuals (or strains) used for genetic analysis. When the density is low, this results in wide QTL peaks with large confidence intervals. Much better resolution, down to the gene level, can be reached by analyzing populations (like the human population for example) that have accumulated huge densities of recombination breakpoints over many generations of random crosses. This observation led a group of geneticists of The Jackson Laboratory to develop the Diversity Outbred (DO) stock, by continued random mating of 144 partially inbred lines of the Collaborative Cross. Each mouse of this stock is genetically unique, and once genotyped by using high-density genotyping arrays (Li et al. 2005; Churchill et al. 2012), it allows unparalleled resolution for QTL mapping. Groups of DO mice approximate the genetic diversity and level of heterozygosity found in human populations (i. e. an average of 390 recombination events per genome at G10) and can be used to validate previously identified QTLs. Groups of DO mice approximate the genetic diversity and level of heterozygosity found in human populations.

10.13 Cloning QTLs

Once a QTL has been identified, confirmed and assigned to a small chromosomal region, identifying the quantitative trait gene (QTG) responsible for the effect observed is the ultimate goal. Even though substantial progress in the knowledge of the mouse genomic sequence has been made in recent years (see Chap. 5), this last step remains a difficult enterprise. There is no unique strategy to go from a genomic region to the gene and it is generally a combination of approaches that will provide clues which, confronted and interconnected, will point at candidate genes eventually submitted to functional analysis.

When the QTL location has been narrowed to a region of a few Mb using congenic and sub-congenic strains, which may require the production and phenotyping of large numbers of animals, the strategies for identifying the causative gene resemble those used for Mendelian traits. They include, in particular, the comparison of whole-genome or whole-exome sequences, the production and in silico analysis of gene expression data, and thorough literature review and database searching to collect detailed information on gene function. One should also carefully look for data coming from other animal models or human conditions. These investigations should lead to a limited number of candidate genes that must be submitted to functional evaluation. The most appropriate testing depends on the nature of the trait and the phenotype that best characterizes the QTL effect.

10.13.1 Analyzing the DNA Polymorphisms in the QTL Region

When the interval is significantly reduced (i.e. less than 1 or 2 Mb), which may require breeding, phenotyping, and genotyping thousands of mice, it is then possible to look at the genome structure focusing on genetic issues. The first thing to establish, when possible, is an exhaustive list of the genes (10–30 on average) which map within the interval, with the likely function of each of them, when this is known from genome annotation. Nowadays, this step of QTL analysis is made somewhat easier if we remember that the genome of several inbred strains has been completely sequenced, making the alignments between the parental strains easier and faster. While comparing these alignments, it is important to check for the possible existence of indels and more generally the integrity of the different genes in the two parental strains. Small-sized deletions and insertions are common findings in the mammalian genome and even though many of them exhibit no clear effect on the phenotype when homozygous they may nevertheless entail slight phenotypic variations.

Gene copy number variations (CNVs) are also important structural differences, which may account for quantitative phenotypic differences (see Chap. 5 for comments; Cutler and Kassner 2008). Finally, SNPs are very important structural variations to look at for two main reasons. (i) First, because among the most recently published results reporting the successful positional cloning of a QTL (whatever the species) a majority indicate that SNP differences have been the starting point, with one of the non-synonymous SNPs being associated with a conformational change often leading to a difference in activity of the encoded protein. (ii) SNPs are also important polymorphisms to look at because they can help in the determination of the ancestral origin of the haplotype containing the QTL under investigation and accordingly can suggest comparisons to be made with strains unrelated to the parental strains but segregating for the same QTL. Any SNP that might be causative of a missense mutation or a splicing defect would require special attention. Nonsense mutations, generating null or hypomorphic allele, are candidates for qualitative mutations but have not been often recognized as being responsible for quantitative phenotypic differences.

Based on the information collected in several species (including plants), the genetic alterations that are the best candidates to account for phenotypic differences in quantitative trait inheritance are those that result in proteins slightly modified in their structure, expression level or stability in time but not in loss or gain of function.

SNP analysis is a logical and straightforward approach but it can sometimes be extremely difficult when, for example, the QTL encompasses a SNP-rich region. In this case thousands of SNPs must be analyzed, with many of them being irrelevant or outside the coding regions.

In conclusion, in many instances the structural variations that can be observed at the sequence level are insufficient to provide an answer in terms of gene identification. Other investigations are necessary to unravel the biology of the candidate genes: for example where, when, and at what level they are expressed.

10.13.2 Quantitative Complementation

Validation of a candidate gene is generally achieved by performing trans-complementation with another allelic form of the candidate gene, which can be either a mutant allele that already exists somewhere in a genetic repository or can be engineered especially for this purpose.

A strategy that seems to be efficient, when applicable, is known as *QTL-knockout interaction* (Darvasi 1998). The experimental protocol requires no less than four strains and four crosses. Strain A, encoding for the "high" allelic form at the QTL (Q^h) in question is mated with a strain carrying a null (knockout) allele of the gene of interest (m^-) and with the co-isogenic strain carrying the wild-type allele of the same gene (m^+). Similarly, two other crosses are made with the strain B encoding the "low" allelic form (Q^l) and the same two strains as above, i.e. (m^-) and (m^+). A greater phenotypic difference between the (Q^h)/(m^-) and (Q^l)/(m^-) genotypes than between the (Q^h)/(m^+) and (Q^l)/(m^+) genotypes provide some evidence of quantitative failure of the mutation to complement the QTL alleles and validate the candidate gene. Theoretically, the QTL/knockout interaction test requires the use of strains co-isogenic for the knockout (m^-) and wild type (m^+) alleles to avoid potentially confusing interactions of other genes in the background.

In summary, although only a few dozen rodent QTLs have been cloned to date, it appears from studies conducted in other species that QTLs are generally made out of non-null allelic variants. These variants are generally characterized by differences in gene expression level or by subtle structural variations that translate into differences in terms of activity for the encoded proteins. This is a major difference from qualitative or Mendelian traits, where null mutations are quite common.

Thousands of QTLs have been mapped onto the mouse genome during the last two decades. However, only a few genes underlying complex traits have been successfully identified, and fine mapping of QTL genes still remains a challenge for mouse geneticists.

10.14 The Analysis of Expression QTLs (eQTLs)

The analysis of gene expression, for example by using expression arrays or RNA sequencing, allows the discovery of quantitative differences, sometimes important, between strains or individuals. Gene expression level is a quantitative phenotype, controlled by *expression QTLs* (eQTLs), amenable to QTL mapping using the methodologies described for other phenotypes.

A number of published studies have shown that most eQTLs are located in *cis*, i.e. in the vicinity of the expressed gene. They most likely correspond to classical regulatory elements such as promoters, enhancer, 3′UTRs, etc (see Chap. 5). In this case, one eQTL influences the level of expression of a single gene. However, a small fraction of eQTLs appear to control the expression of multiple genes

located on different chromosomes. Although the resolution of eQTL mapping cannot rule out the possibility that these genes are controlled by several, independent but tightly linked, regulatory factors, it is hypothesized that they correspond to key elements of regulatory networks.

Notably, RIS have proved very valuable in mouse studies, in particular the BXD set established between C57BL/6J and DBA/2J strains, which currently comprises around a hundred strains. Gene expression levels can be measured and compared in every inbred strain comprising a given set of multiple individuals, in males and females, in multiple tissues and organs, at different ages and under multiple conditions. Given that these strains have been extensively genotyped, this experimental design allows quick identification of eQTLs, either sex-, age- or tissue-specific. It also provides optimal power to identify correlations between the expression levels of different genes or between organs across a series of genetically different inbred strains. Most interestingly, the BXD set has been heavily investigated for a wide variety of phenotypes including metabolic, behavioral, immunological, and many other traits. It is now possible to search in silico for associations between phenotypic data and gene expression levels, and identify, at a genome-wide scale, genes whose expression correlates, positively or negatively, with a trait of interest. This approach, which will certainly reveal unsuspected relationships, has been made possible and publicly available by the development of centralized databases and analytical tools, in particular by GeneNetwork (http://www.genenetwork.org).

10.15 The Case of Modifier Genes

When studying the phenotype of mouse mutations, variations in phenotypic expression are common observations. Whereas, for example, mice homozygous for the Tyr^{c} allele are always completely albino, mice homozygous for the same $Lepr^{db\text{-}Pas}$ mutation (leptin receptor non-functional allele) have a phenotype that is very different when the mutant allele is on the PWK/Pas background or the DW/J background. In the former case, the mice are moderately fat but mostly diabetic and secrete enormous amounts of urine (sometimes up to 50 ml per day!). On the DW/J background, on the contrary, the mice carrying the same allele grow to become very fat (up to 80 g) but urinate almost a normal quantity. This observation of great variations in the phenotypic expression is more the rule than the exception in mammals and for some mutations the situation is sometimes extreme. Many dominant bone mutations have probably been lost because they could not be transmitted from one generation to the next for a lack of expressivity while others were so severely expressed that they appeared to be incompatible with life.

This aspect of background-dependent phenotypic variation also exists in human populations where it has a great importance making a specific syndrome more or less compatible with every-day life. The same situation is also common with certain forms of cancer with a clear genetic determinism and great variations in terms

of severity depending on the affected person. Colon cancers triggered by mutant alleles at the *APC* gene are a good example, as are the deficiencies in the enzyme ferrochelatase (FECH) producing a syndrome with extreme variations depending on the patient.

All the variations reported above are the consequence of multiple genetic interactions between the mutant gene and a number of unknown genes modulating the expressivity of the phenotype. The effect of these modifier genes can be measured by quantitative variables and the genes can be identified using the general strategies described above. This research is extremely difficult in humans but the mouse can provide relevant models. The genes identified in mice may provide at least an indication of metabolic pathways and physiological processes that have a significant probability of being conserved between the two species. Interesting and promising results have been obtained recently.

10.16 Conclusions

The susceptibility of humans and domestic animals to certain forms of cancer or to most infectious diseases, just like the variations in expression of some metabolic diseases, is often influenced by genetic factors. To date, the nature of these factors, although obvious, is unknown and this is unfortunate because it might be used as a way of influencing the prognosis of many diseases. For this reason, one can predict that studies on the genetic determinism of complex traits will expand in the years to come. In such studies, the mouse will undoubtedly play a pivotal role because in this species, more than in any other, powerful tools and strategies are available. This will certainly help researchers to know better the composition of the QTLs at the molecular level.

References

Brahic M, Bureau JF (1998) Genetics of susceptibility to Theiler's virus infection. BioEssays 20:627–633

Burgio G, Szatanik M, Guénet JL, Arnau MR, Panthier JJ, Montagutelli X (2007) Interspecific recombinant congenic strains between C57BL/6 and mice of the Mus spretus species: a powerful tool to dissect genetic control of complex traits. Genetics 177:2321–2333

Burgio G, Baylac M, Heyer E, Montagutelli X (2012) Exploration of the genetic organization of morphological modularity on the mouse mandible using a set of interspecific recombinant congenic strains between C57BL/6 and mice of the *Mus spretus* species. G3 (Bethesda) 2:1257–1268. doi: 10.1534/g3.112.003285

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ,

Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpää MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, Tarantino LM, Threadgill D, Toth LA, Valdar W, de Villena FP, Warden C, Whatley S, Williams RW, Wiltshire T, Yi N, Zhang D, Zhang M, Zou F; Complex Trait Consortium (2004) The collaborative cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137

Churchill GA, Gatti DM, Munger SC, Svenson KL (2012) The diversity outbred mouse population. Mamm Genome 23:713–718

Cutler G, Kassner PD (2008) Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. Cytogenet Genome Res 23:297–306

Darvasi A (1998) Experimental strategies for the genetic dissection of complex traits in animal models. Nat Genet 18:19–24

Grisel JE, Belknap JK, O'Toole LA, Helms ML, Wenger CD, Crabbe JC (1997) Quantitative trait loci affecting methamphetamine responses in BXD recombinant inbred mouse strains. J Neurosci 17:745–754

Groot PC, Moen CJ, Dietrich W, Stoye JP, Lander ES, Demant P (1992) The recombinant congenic strains for analysis of multigenic traits: genetic composition. FASEB J 10:2826–2835

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Li R, Lyons MA, Wittenburg H, Paigen B, Churchill GA (2005) Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics 169:1699–1709

Mackay TFC (2009) Q&A: genetic analysis of quantitative traits. J Biol 8:23

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565–577

Singer JB, Hill AE, Burrage LC, Olszens KR, Song J, Justice M, O'Brien WE, Conti DV, Witte JS, Lander ES, Nadeau JH (2004) Genetic dissection of complex traits with chromosome substitution strains of mice. Science 304:445–448

Threadgill DW, Hunter KW, Williams RW (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. Mamm Genome 13:175–178

Threadgill DW, Churchill GA (2012) Ten years of the collaborative cross. G3 (Bethesda) 2:153–156

Xiao J, Liang Y, Li K, Zhou Y, Cai W, Zhou Y, Zhao Y, Xing Z, Chen G, Jin L (2010) A novel strategy for genetic dissection of complex traits: the population of specific chromosome substitution strains from laboratory and wild mice. Mamm Genome 7–8:370–376

Publications of General Interest

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow

Flint J, Valdar W, Shifman S, Mott R (2005) Strategies for mapping and cloning quantitative traits in rodents. Nat Rev Genet 6:271–286

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, MA Sinauer

Mackay TF (2001) The genetic architecture of quantitative traits. Ann Rev Genet 35:303–339

Xu S, Atchley WR (1996) Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143:1417–1424

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468

Computer softwares

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 35:124–129

GeneNetwork: http://www.genenetwork.org

Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421–1428

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387

Glossary

5'-untranslated region (5'-UTR) or leader sequence A sequence measuring 100 to several thousands bp, between the transcription initiation site and the initial AUG. This sequence usually contains a ribosome binding site (RBS), known as the Kozak sequence (gcc)gcc(A/G)ccAUGG, which includes the AUG initiation codon.

3'-untranslated region (3'-UTR) or trailer sequence A sequence, downstream of the stop codon, which is required for the processing of mRNAs. The polyadenylation signal, composed of sequences like AAUAAA, is an important component of the 3'-UTR.

Acrocentric A chromosome in which the centromere is located near one end. The 40 chromosomes (19 pairs of autosomes, X and Y) of the normal laboratory mouse are all acrocentric (see Chap. 3).

Additive inheritance A situation where the observed phenotype results from the cumulative, individual effects of the alleles carried at a group of loci, with no epistatic interactions and no allele-dependent effect of the environment.

Allele (allelomorph) One of the alternative forms of a gene, which may differ from the others by a variety of polymorphisms ranging from a single nucleotide (an SNP) to a sequence of several nucleotides.

Alternative splicing This expression means that not all exons in a gene are assembled to form the maturated mRNA, on the contrary, some are deleted (or spliced out). It is estimated that ~95% of multiexonic genes of the mammalian genomes are alternatively spliced. However, not all genes have introns.

Anchor locus A locus whose location in the genome is very precisely known, used as a landmark in the construction of genetic maps. A gene that is, at the same time, characterized phenotypically and at the DNA level (sequence) and that has orthologs in several other species is an ideal anchor locus.

J.L. Guénet et al., *Genetics of the Mouse*, DOI 10.1007/978-3-662-44287-6

Aneuploidy Any variations from the normal 2n (euploid) chromosome number. Triploidy (3n), haploidy (n), trisomy (2n+1), monosomy (2n−1) are classical aneuploidies. Aneuploidies sometimes involve only a portion of chromosome.

Annotation Gene annotation consists of establishing the structure and the likely function(s) of a given gene or DNA sequence.

Anonymous locus A sequence of DNA with no known function but with at least two allelic forms that can be followed generation after generation through some form of DNA analysis.

Analysis of variance (ANOVA) ANOVA is a collection of statistical methods, developed by R.A. Fisher, and used to analyze the robustness of observed differences between groups of individuals. Using ANOVA, a statistically significant result would be when a probability (*p*-value) is less than a threshold (significance level), and justifies the rejection of the null hypothesis (the observed difference between two samples may result from chance only) (see chi-square (χ^2) test).

Anti-sense mRNA An RNA molecule or oligonucleotide that is complementary to an mRNA molecule and can form a duplex with it. Antisense mRNAs interfere with mRNAs translation (in mammals) and provoke their destruction by ribonucleases specific for double-stranded molecules.

Autosome Any chromosome other than the X or Y chromosome. The normal mouse karyotype consists of 19 pairs of autosomes.

B1/B2 repeats Also designated Short Interspersed Nuclear Elements or SINE (see Chap. 5). B1/B2 are the two most prominent classes of repetitive elements in mammalian genomes, with a size of 500/600bp. The mouse B2 repeats are equivalent to the human *Alu* sequences.

BAC (Bacterial artificial chromosome) A cloning vector that uses the origin of replication of the functional fertility plasmid (F plasmid) of *E. coli*. BACs can take inserts with a size up to 150–350 kb and can be used for the production of transgenic mice. BACs are more reliable vectors than YACs.

Backcross Literally, a backcross is the cross of a hybrid with one of its parents or an individual genetically similar to it. In practice, it is a cross between an animal heterozygous for two different alleles and an animal homozygous for one of the two alleles in the heterozygous parent (for example *A/a* × *a/a* or *A/a* × A/A). In the particular case where the cross is between a heterozygous mouse and a mouse homozygous for the recessive allele (*A/a* × *a/a*), the backcross is called a testcross (see Chaps. 2 and 4).

BLAST (basic local alignment of sequences tool) A popular computer program that searches for similarity of a selected sequence to sequences stored in a database.

Blastocoel(e) or blastocele The fluid-filled cavity of a blastocyst.

Bin A bin is a group of syntenic genetic markers that have not been separated (ordered) by meiotic recombination in a given cross (see syntenic).

CAGE (Cap-analysis of gene expression) A technique based on the preparation and sequencing of concatamers of DNA tags deriving from the initial 20 nucleotides from the 5′ end of mRNAs allowing high-throughout analysis of the capped transcripts population in a biological sample. CAGE detects the transcriptional activity of each promoter.

Candidate gene A gene thought to be likely responsible for an observed phenotype because of its genetic localization and/or likely function. Candidate genes are, in most instances, suggested by positional cloning or association studies.

CAT (or CCAAT, or CAAT)-box A consensus sequence GGCCAATCT, inserted 75−80 bp upstream from the transcription start site. Some genes with relatively ubiquitous expression do not have this GGCCAATCT sequence.

Centimorgan (cM) The unit used to describe genetic distances. A centimorgan is the distance between two genes that will recombine with a frequency of exactly 1% (see Chap. 4). Genetic distances are additive, while recombination fractions are not. Genetic distances are computed from recombination fraction using a mapping function.

Centromere The centromere is the part of a chromosome that links sister chromatids and attaches to the spindle fibers (through the kinetochore) (see Chap. 3).

Chi-square A statistical test developed by Karl Pearson and commonly used by geneticists to compare observed data from a given experiment (for example, the different proportions of phenotypes in a progeny) with the theoretical data that would be obtained according to a specific hypothesis. The chi-square (χ^2) test appreciates whether the differences between observed and expected proportions could result from chance (the *null hypothesis*) or if they are more likely to be due to an explanatory factor. (see Chap. 4).

Chiasma (plural: chiasmata) The point where two homologous non-sister chromatids exchange genetic material during meiosis.

Chimera (or Chimaera) An organism that is composed of two (or more) populations of genetically distinct cells originating from two or more embryos. Chimeras generally result from human intervention (see Chap. 2—see also Mosaics).

Chromatin The complex of DNA, RNA and protein that makes up the core of chromosomes. There are two sorts of chromatin: *euchromatin*, whose structure is open, allowing the transcription of the DNA component, and the *heterochromatin*, which is made of repetitive sequences that are not transcribed.

Chromosome banding A staining process that produces a discrete, reproducible pattern of light- and dark bands that can be used to identify individual chromosomes and chromosomal regions. See G-bands and Q-bands (see Chap. 3).

Clone-by-clone A strategy used for sequencing the human genome. The genome in question is cloned into BACs, the clones are ordered and shotgun-sequenced (see shotgun sequencing). Finally, the sequence is assembled by ordering head-to-tail the sequences from adjacent BAC clones.

Coding sequence A stretch of DNA or RNA whose sequence ultimately determines the sequence of a protein (see Chap. 5). The coding sequence excludes introns.

Codominance A kind of allelic interaction in which an animal heterozygous for two alleles (A^1 and A^2) at the A locus, expresses at the same time, the phenotypes that would be observed in the two corresponding homozygotes (A^1/A^1 and A^2/A^2). Codominance is more the rule than the exception in mammals.

Coisogenic A strain of mice that differs from an established inbred strain by a single point mutation at a given locus (see Chap. 9).

Collaborative cross (CC) A panel of recombinant-inbred strains, generated by randomizing the genetic diversity of existing inbred mouse resources from the three major *Mus musculus* subspecies (*M. m. musculus, M. m. domesticus*, and *M. m. castaneus*). A useful tool for mapping multigenic traits (see Chaps. 9 and 10).

Comparative genomic hybridization (CGH) CGH is a molecular method for assessing possible copy number variations (CNVs), through independent labeling of a reference sample and a test sample of denaturated DNA with fluorophores of different colors (usually red and green) (see Chap. 5).

Complementary uniparental disomies/nullosomies A normal, euploid (2n) embryo resulting from the fusion of two aneuploidy gametes. When one parent contributes two chromosomes of the same pair and the other none, this results in an embryo with complementary uniparental disomy/nullosomy (abbreviation UpD or UPD). Some of these embryos are viable, others are not due to the differential imprinting of the chromosomes in the gametes (see Chap. 6).

Complex diseases Diseases whose etiology consists of a mixture of environmental and genetic factors. In many instances the genetic factors are numerous and/or of various "strength".

Compound heterozygote An individual heterozygous for two mutated alleles at a given gene (for example A^{m1}/A^{m2}).

Congenic A strain of mice that is formed by introgressing (i.e. backcrossing repeatedly) a chromosomal segment carrying a locus of interest into an inbred parental strain for ten or more generations (see Chap. 9). For example, B6.C-*Tyr*^{c} is a congenic strain with C57BL/6 (B6) background carrying a segment of chromosome 7 from BALB/c (C) origin that harbors the albino mutation (*Tyr*^{c}), resulting in albino B6 mice.

Conplastic Conplastic strains have the same nuclear genome but different mitochondrial genome. A conplastic strain is developed by transferring the nuclear genome from one inbred strain into the cytoplasm of another (the donor parent

is always the female parent during the backcrossing program). A minimum of 10 backcross generations is required for a strain to be 100% conplastic. Conplastic strains are useful for studying the role of mtDNA.

Consensus sequence A theoretical sequence that represents the nucleotides most often found at each position when a number of different sequences are compared.

Consomic strains Consomic strains are a particular case of congenic strains, in which the transferred segment consists of an entire chromosome. Also called chromosome substitution strains (CSS) (see Chap. 9).

Contig A DNA region represented by a set of (head-to-tail) overlapping genomic clones (BACs, P1 or YACs) that together span a region of the genome larger than that covered by any one clone (see Chaps. 5 and 9).

Copy number variants (or variation) (CNVs) Segments of DNA that are present in an individual with a copy number that is different from the reference genome (C57BL/6).

Cosmid A hybrid plasmid containing the cos site from the lambda phage, used as a cloning vector combining the properties of a plasmid and a phage. Cosmids allow recombinant DNA molecules of 30 to 50 kb to be packaged into phage particles in vitro. Upon infection into a host cell the recombinant DNA molecule replicates as a plasmid.

Coverage The average number of times a same nucleotide is sequenced in genome shotgun sequencing. Higher coverage ($\times 8$ - $\times 10$) provides higher reliability of sequence data.

C_pG island Short DNA sequences (a few kb long) that are GC-rich, C_pG-rich, and predominantly non-methylated. C_pG islands are associated with the 5'-end of genes and most, perhaps all, of these sequences are sites of transcription initiation (see Chap. 5).

Crossing-over A crossing-over is the reciprocal exchange of genetic material between homologous chromosomes during meiosis (see Chap. 4).

Cytogenetics The part of Genetics that deals with chromosome structure, chromosome behavior during meiosis and pathology resulting from chromosomal breakage or imbalance.

Deleterious allele An allele with a more or less severe effect on the phenotype. A missense allele can be a deleterious allele if the substituted amino acid impairs the function of the protein. A null-allele is often deleterious.

Deletion mutations The loss of one or more nucleotides or, sometimes, a fragment of chromosome.

Deme A breeding unit in natural populations of mice. A deme usually consists of one dominant male with up to six-eight females (see Chap. 1).

DGGE (denaturating gel gradient electrophoresis) A sensitive gel-based technique for detecting single nucleotide changes within orthologous or allelic PCR products that have been denaturated and gel-fractionated as single strands.

Dinucleotide Two successive nucleotides on the same DNA strand, written with the 5' nucleotide first.

Disjunction The normal process by which the two homologs of each chromosome in a meiotic cell separate and move to different gametes as a single unit (see Chap. 2).

Distal A relative term meaning closer to the telomere; the opposite of proximal.

DNA marker A short sequence of DNA with allelic variation that can be followed directly by a DNA-based assay such as hybridization techniques, PCR or sequencing (see Chap. 2).

DNA methylation The addition of a methyl group to cytosine to form 5-methylcytosine in CpG dinucleotides. Between 8 and 10% of the cytosine nucleotides in the mouse genome are methylated. Methylation alters gene expression and plays a major role in parental genomic imprinting.

Domain Part of a protein where the polypeptide chain folds to form a discrete globular structure that has a defined function. Also apply to the sequence of DNA encoding the protein domain in question.

Dominant allele Dominance describes the relationship of one allele to a second at the same locus when an animal heterozygous for these alleles expresses the same phenotype as an animal homozygous for the first allele. The second allele of the pair is considered recessive. Mice homozygous for the agouti allele (*A*/*A*) as well as mice heterozygous for the agouti and nonagouti alleles (*A*/*a*) carry hairs with yellow pigment (phaeomelanin) and black pigments (eumelanin) and cannot be distinguished, while nonagouti (*a*/*a*) mice have black hairs (eumelanin). Hence *A* is dominant, whereas *a* is recessive. True dominant alleles (such as *Caracul—Krt71*^{Ca} or *Rex—Krt25*^{Re}) are uncommon in the mouse. In many cases mice homozygous for the dominant allele are recessive lethal (A^y/+ mice are yellow while A^y/A^y genotypes die in utero).

Dominant negative mutation A mutation that prevents a wild-type allele in the same cell from functioning. Dominant negative mutation commonly acts by producing an altered polypeptide that prevents the assembly of a multimeric protein. A number of mutations at the *Kit* locus, that encodes the KIT tyrosine kinase receptor act in a dominant-negative manner. Mutations in the gene *Col26a1*—encoding collagen–are dominant negative.

Double heterozygotes Animals heterozygous at two loci on the same chromosome. Depending on the alleles assortment, double heterozygotes can be in coupling (*AB*/*ab*) or in repulsion (*Ab*/*aB*).

Downstream A region of the DNA molecule that lies 3' to the point of reference.

Draft sequence Preliminary form of a sequence with gaps and errors.

Electrophoretic variant Allelic variant of an enzyme (allozyme) with altered electrophoretic mobility due to a charge-changing amino-acid substitution.

Embryonic stem (ES) cells Pluripotent cells taken from the inner cell mass of blastocyst stage embryos, which are cultivated in vitro as a permanent cell line. These cells can be genetically modified in a number of ways and, once reinserted into a developing embryo (blastocyst), they are capable of participating in the formation of the germline (see Chap. 8).

Embryonal carcinoma EK cells Pluripotential cells derived from spontaneous or experimental teratocarcinomas (EC) or from embryos (EK cells) that were used for studying some aspects of early mouse embryogenesis. These cells were "precursors" of ES cells. They were named after Evans and Kaufman, who made extensive use of them.

ENCODE The ENCODE project (ENCyclopedia Of DNA Elements) has set as its major aim to establish a catalog of all the structural and functional elements of the genome.

Endogenous retroviruses (ERVs) A DNA sequence resulting from the integration of more or less complete retroviral copies into the mouse genome. ERVs are flanked by two long terminal repeats (LTR).

Endonuclease An enzyme that, unlike exonucleases, cleaves RNA or DNA molecules at an internal position rather than progressively from either the 5' or 3' end.

Enhancer A (50–1500 bp) DNA sequence that increases the expression of a gene when bound to a transcription factor (activator). Enhancers are generally cis-acting and in most cases they exert their influence irrespective of their location or orientation. Enhancers are numerous in mammalian genomes.

Ensembl Genome browser and database of sequenced genomes jointly maintained by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute. Accessible on the web by the URL www.ensembl.org.

Epistasis Epistasis characterizes the interaction between non-allelic genes in which one gene suppresses or enhances the expression of another. When homozygous the albino allele (Tyr^c) at the *Tyr* locus is epistatic over all other genes with an action on coat color. Epistasis more generally describes cases where the phenotype controlled by two loci cannot be predicted by considering the two loci independently. Epistasis is a very common situation among genes that control complex traits.

Estrous cycle The estrous cycle consists of cyclic physiological changes induced by reproductive hormones in mammalian females. In the mouse, the estrous cycle lasts 4 to 6 days and is arbitrarily divided into four stages: proestrus, estrus, metestrus, and diestrus.

Ethyl methane sulfonate (EMS) A potent alkylating agent that is active on postmeiotic germ cells and ES cells.

Ethyl nitrosourea (ENU) A highly potent alkylating agent used to introduce random mutations (mostly base pair changes) in the mouse DNA. ENU is active on pre- and post-meiotic germ cells.

Euchromatin The main fraction of chromosomal DNA that is uncoiled during interphase, and contains transcriptionally active regions. The other fraction is *heterochromatin*.

Exon trapping Special technique used in the past to search for coding sequences (exons).

Exon The part of a gene sequence that remains present within the messenger RNA (mRNA) after introns have been removed by RNA splicing. The word was coined from "expressed region".

Exonuclease An enzyme that, unlike endonucleases, degrades progressively RNA or DNA molecules from either the 5' or 3' end rather than at an internaI position.

Expressed sequence tag (EST) ESTs are short sub-sequences (~350 to 500 bp) of a cDNA sequence, starting in general from the 3' end, sometimes from the 5' end. ESTs can be used as molecular probes to retrieve the complete transcript of a gene.

Expressivity A genotype exhibits variable expressivity when individuals with that genotype differ in the extent to which they express the phenotype normally associated with that genotype. Mice heterozygous for the brachyury mutation (*T*/+) are usually characterized by short tails, but the length of their tail is highly variable from one mouse to the other. The *T* mutation exhibits variable expressivity. Such variations can be caused by environmental factors, by modifier genes or by chance (developmental noise).

F1 The offspring of a cross between two different inbred strains (see also hybrid F1).

FANTOM research project Functional Annotation of the Mouse Genome (FANTOM) is an international research consortium founded in 2000 by Dr. Hayashizaki and his colleagues at RIKEN in Tokyo, Japan with the aim to functionally annotate the mouse DNA sequence. FANTOM has since developed and expanded over time to encompass the regulation of genes, networks of genes and their impact in disease.

Fingerprinting Any method that identifies unique features of a clone that can be used to determine overlaps between this clone and other clones in a library. Restriction sites are useful tools for DNA fingerprinting.

Finished sequence The final form of a sequence from the Mouse Genome containing less than 1 error in 10,000 bp.

FISH Fluorescent in situ hybridization (see Chap. 4).

Fluorophore A chemical substance that can re-emit light upon stimulation by light of a particular wavelength (excitation light).

Forward genetics (positional cloning) A strategy whose aim is to characterize the structural alteration(s) at the genome level that is (are) associated with (or responsible for) a specific phenotype. The strategy proceeds from phenotype to genotype and, for this reason, it is often referred to as forward genetics. It is the opposite of reverse genetics.

Frameshift mutation Deletion or insertion of a few base pairs (not a multiple of 3!) that alters the reading frame downstream of it.

Functional genomics The study of the function of genes in a genome.

Genetic drift The unavoidable evolution of the genetic structure of a population over generations. In fully inbred strains, genetic drift results from spontaneous neutral mutations that disappear or become fixed throughout generations. Residual heterozygosity in partially inbred strains is often responsible for genetic drift.

Genetic marker Any gene or short DNA sequence whose chromosomal location is precisely known and whose structure exhibits variations among the individuals of the same strain or species. A genetic marker can be a phenotypic marker or molecular marker (see Chap. 4).

Genome The total genetic information present within a single cell nucleus of an animal. The haploid genome of the mouse is 3×10^9 bp and encodes 41,968 genes (genes with nucleotide sequence data—as of November 2014) (see Chap. 5).

Genomic Library A collection of DNA clones large enough to guarantee that any sequence of interest in the genome is likely to be present in at least one clone.

Genotype The set of alleles present at one or more loci of a given individual.

Giemsa A stain used to accentuate visually the difference between bands and interbands on metaphase chromosomes (see Chap. 3).

Golden path The set of minimally overlapping cloned DNA that covers a large segment (and ideally all) of the genomic DNA of a given chromosome.

Haldane's rule In interspecific hybrids, the heterogametic sex is more severely affected by traits that concern viability or sterility. For example, hybrids between *Mus spretus* and *Mus musculus domesticus* are male sterile but female fertile.

Haplotype Pertaining to a particular set of alleles that are found together at linked loci. In linkage studies, haplotypes provide very reliable data for determining the order of loci (see Chaps. 4 and 9).

HAVANA (for Human and Vertebrate Analysis and Annotation of the genome): A program undertaken by a team at the Sanger Institute for the systematic and careful annotation of the human, mouse and zebrafish genomes.

Hemizygous Describes an individual who has only one member of a chromosome pair or chromosome segment rather than the usual two. X-linked genes in males are hemizygous. Chromosomal deletions and transgenic insertions are often hemizygous.

Heterozygote An animal with two different alleles at a particular locus. In this case, the locus is considered heterozygous.

Hierarchical shotgun sequencing (HSS) A sequencing strategy that makes use of cloned DNA with large inserts previously assembled into a series of overlapping contigs.

High-resolution/high density map A genetic map with a great number of genetic (mostly molecular/DNA) markers accurately mapped. High resolution and high density maps are constructed based on large sized progenies of crosses between strains with a high level of genetic polymorphism. Such maps have been instrumental for the complete sequencing of the mouse genome.

Histocompatible Pertaining to a genetic state in which cells from two animals can be cross-transplanted without triggering rejection. Histocompatibility is controlled by many genes. *H2*, on chromosome 17, is the major histocompatibility complex or MHC.

Histones A class of alkaline proteins whose function is to package (protect ? isolate ?) the nuclear DNA. Histones are the major components of chromatin.

Homolog A gene that shares with another gene a common ancestry. The term homolog may apply to the relationship between genes separated by the event of speciation (see ortholog) or genetic duplication (see paralog).

Homozygote An animal with two identical alleles at a particular locus.

Hotspot, recombinational A region of chromosome, usually less than a few kilobases in length, that participates in crossover events at a very high rate relative to neighboring "cold" regions of chromosome (see Chap. 4).

Hybrid F1 The offspring of two homozygous individuals (e.g., inbred strains). For example (C57BL/6 x C3H)F1 mice come from a cross between a female C57BL/6 and a male C3H.

IBD/S Identical by Descent/State two genes or DNA segments with identical nucleotide sequences in two or more individuals are said to be *identical by state* (IBS). If they are identical because they were inherited from a common ancestor then they are said to be *identical by descent* (IBD).

Ideogram A schematic representation of chromosomes indicating their relative size, the position of the centromere and their banding patterns.

Imprinting A genetic mark that alters the expression of a gene. Imprinting varies depending on the parent a given gene is inherited from. Only a small subset of

genes in the mammalian genome is imprinted. The biological meaning of imprinting is not yet known but abnormal imprinting can result in pathology.

Inbred Strain A strain that is essentially homozygous at all loci, typically produced by brother-sister matings for at least 20 successive generations. BALB/c and C57BL/6J are popular mouse inbred strains.

Intercross A cross between two identical hybrid individuals (*A*/*a* x *A*/*a*).

Interference When a crossover occurs in a region, this affects the likelihood that another crossover event occurs in the adjacent region. This interaction is called interference. Interference varies in intensity among species.

Intra Cytoplasmic Sperm Injection (ICSI) (or micro-insemination) A procedure that consists in the injection of a sperm head into the cytoplasm of the oocyte in general by using a piezo-driven micromanipulator (see Chap. 2).

Introgression The introduction of one or several alleles of foreign origin into a different gene pool. The word introgression is sometimes used when a segment of chromosome containing a gene of interest is selectively transferred by sexual reproduction, from its original background into another strain. Artificial introgression of a given gene, from one strain into another, results in a strain congenic for the gene in question. Introgression of a complete chromosome from a donor strain into a recipient strain results in a consomic strain.

Introns Introns correspond to the nucleotide sequences within a gene that are removed by RNA splicing while the final mature RNA product is being generated. The word was coined from "intragenic region". Some mammalian genes, such as those encoding histones or tRNAs, have no introns. Some introns contain sequences that are transcribed in non-protein coding RNAs.

Isogenic Individuals sharing identical genes (alleles). Identical twins, clones, and individuals from an inbred strain are isogenic. Inbred strains are isogenic and homozygous at all of their loci.

Junk DNA Coined by the geneticist Susumo Ohno, this expression referred to the non protein-coding fraction of genomic DNA. Nowadays, geneticists consider that the proportion of "junk DNA" in a mammalian genome is limited to only a few percent and while most of the genomic DNA is transcribed (see Chap. 5).

Karyotype The number and appearance of the chromosomes in a eukaryotic cell.

Kilobase A stretch of 1,000 base pairs of DNA (abbreviation: kb).

Knock-in The targeted insertion of a (cloned) exogenous gene into the mouse genome with the aim to disrupt an endogenous gene while expressing the transgenic one.

Knockout (KO) An animal with one of its gene inactivated by genetic engineering. A knockout gene can also result from a knock-in (see Chap. 8).

Kozak sequence (gcc)gcc(A/G)ccAUGG (see Chap. 5).

Linkage group A set of loci in which all members are linked either directly or indirectly to all other members of the set. A linkage group is equivalent to the genetic information associated to a single chromosome.

Linkage map Genetic or meiotic map. A linkage map is based on linkage data.

Linkage Pertaining to the situation where two loci are close enough to each other on the same chromosome that recombination frequency between them is reduced to a level significantly less than 50%.

Locus Any genomic site, whether functional or not, that can be mapped through formal genetic analysis.

LOD score The logarithm (base 10) of odds. The lod score is a statistical test developed by Newton E. Morton that is used in linkage analysis. It compares the likelihood of obtaining the test data if the two loci were indeed linked, to the likelihood of observing the same data in the absence of linkage. A LOD score of 3 or more is traditionally considered significant to confirm linkage between two loci.

Long Conserved Non-Coding Sequences (see Ultraconserved Elements - UCE)

Long Interspersed Nuclear Elements (LINE) An important category of transposable elements in mammals. Among these LINEs, the L1 family is the most frequent (17–20% of mouse genomic DNA).

Mapping function A mathematical function that converts non additive recombination fractions (because of multiple crossing-overs) into additive genetic distances. Several mapping functions have been proposed (Haldane, Kosambi, etc…) to account for various levels of interference.

Megabase A stretch of 1,000,000 base pairs of DNA (abbreviation: Mb).

MegaMUGA (Mega Mouse Universal Genotyping Array) A genotyping tool involving more than 77,000 informative SNP markers and covering the whole mouse genome with an average spacing of 33 kb between markers. For information concerning Mega MUGA refer to: http://csbio.unc.edu/CCstatus/Media/MegaMUGAFlyer.pdf

Meiosis The process by which diploid germ cell precursors segregate their chromosomes into the haploid nuclei of the gametes.

Meiotic product A single haploid genome within an egg or sperm cell.

Mendelian inheritance/proportions A pattern of segregation for a given phenotype that is (statistically) in agreement with Mendel's laws of inheritance.

Metacentric A chromosome in which the centromere is in the middle of the structure and the two arms of roughly the same size. When the centromere is

shifted towards one end, the word sub-metacentric is used. Sub-metacentric chromosomes have a long arm (symbol *q*) and a short arm (symbol *p*).

MicroRNA or miRNA A short sized (21–25 nt long) single stranded, non-coding RNA molecule which functions in RNA silencing and post-transcriptional regulation of gene expression (see Chap. 5).

Microsatellites A very short unit sequence of DNA (2–6 bp) that is repeated multiple times in tandem. Microsatellites (also called simple sequence repeats or SSRs) are highly polymorphic and have been very useful in linkage analysis (see Chaps. 4 and 5). A polymorphism at a microsatellite locus is also referred to as a simple sequence length polymorphism (SSLP) or Short Tandem Repeat (STR).

Minisatellites A highly polymorphic type of locus containing tandemly repeated sequences having a unit length of 10–40 bp. Minisatellite polymorphisms can be assessed by restriction fragment length polymorphism (RFLP) analysis or by polymerase chain reaction (PCR). Also referred to as variable number of tandem repeat (VNTR) loci (see Chap. 5). These sequences are the base of the original "DNA Fingerprinting" used in forensics.

Missense mutation A non-synonymous substitution in a codon that results in the substitution of an amino acid for another (see Chap. 7). The Eiche's dominant spotting mutation at the *Kit* locus ($Kit^{W\text{-}ei}$) results from the replacement of the Gly amino acid at position 597 by an Ala residue in the KIT receptor kinase receptor.

Model organism Any organism with a phenotype reminiscent of, or similar to a human phenotype. Some mutant genotypes of the mouse are faithful (homologous) models of human diseases, others are much less faithful (analogous). Both models are useful.

Monobrachial homology A mouse heterozygous for two Robertsonian translocations of different origins with one arm in common. For example *Rb(16.17)* and *Rb(5.17)*, are said to be heterozygous with monobrachial homology for chromosome 17.

Monosomic A karyotype with 2n-1 chromosomes. Monosomy can be primary, when one complete chromosome is missing or tertiary if only a fragment of chromosome is missing.

Mosaics Mosaics are organisms composed of cells with a different genetic constitution, although deriving from one and a single conceptus (see Chap. 2). Because one of their two X-chromosomes is randomly inactivated, mammalian females heterozygous for different X-linked alleles, are mosaics.

Mouse Clinic Large-scale phenotyping platforms where mouse mutants or strains are thoroughly analyzed for the greatest possible number of parameters using a panel of highly standardized protocols.

mtDNA Mitochondrial DNA (see Chap. 5).

Multifactorial A trait controlled by at least two factors, which may be genetic or environmental (see Chap. 10). Behavioral differences between inbred strains, such as anxiety, are multifactorial traits.

Mutant allele A mutant allele at a locus is associated with a phenotype distinct from that observed in individuals carrying the most common, so-called wild-type, allele. Non-mutant alleles are often designated wild-type allele i.e. the most common form present at a given locus.

Mutation A new allele that arose abruptly and is present in the genome of an animal but not in the genome of either of its parent(see Chap. 7).

N_2, N_3, N_4 etc. Symbols used to describe the generation of backcrossing and the offspring that derive from it. The N_2 generation describes offspring from the initial cross between an interstrain FI hybrid and one of the parental inbred strains. Each following backcross generation is numbered in sequence (see Chap. 9).

Neutral allele An allele with no noticeable effect on the phenotype. A missense allele can be neutral if the change in nucleotide sequence does not affect the amino acid sequence, or if the amino acid substitution has no effect on the protein function or stability.

Non-coding RNAs RNA molecules that are transcribed from the genome and do not encode protein sequences. The Encyclopedia of DNA Elements (ENCODE) project suggested that over 80% of the DNA in the mammalian genome is transcribed and have an important biological function even if the function in question is not yet elucidated.

Non-disjunction An accident occurring during the meiotic process leading to an abnormal distribution of the chromosomes in the daughter cells (see Chap. 3).

Non-sense mutation The mutation of any codon towards a stop codon. Such a mutation can truncate the protein.

Oligo-nucleotide A chain of nucleotides (nt) usually 10 to 500 nt long. Oligonucleotides are often used as primers for polymerase chain reaction (PCR) amplification.

ORF—Open reading frame The part of a (protein coding) DNA sequence that contains no stop codons.

Orthologs Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Orthologous genes in general retain the same function in the course of evolution. Identification of orthologs is instrumental for reliable prediction of gene function in newly sequenced genomes.

Outcross A cross between genetically unrelated animals.

Overdominance A rrare condition in which the heterozygotes (*M*/*m*) have a phenotype that is more pronounced than that of either homozygotes (*M*/*M* and *m*/*m*)

(see Chap. 6). Mice homozygous for the Mpl^{hlb219} mutation in the thrombopoietin (TPO) receptor MPL (Cys → Arg) have a 80% decrease in the number of platelets in comparison to the wild-type mice. However, mice heterozygous for the same Mpl^{hlb219} allele show an overdominance effect with a significant increase in platelet number.

***p*-arm** The short arm of a sub-metacentric chromosome ("*p*" stands for petit—small in French).

Paralog Paralogs are genes related by duplication within a genome. While orthologs retain the same function in the course of evolution, paralogs evolve new functions, even if these are related to the original one. The Keratin (*Krt*) and Homeobox (*Hox*) genes have many paralogs in the mouse.

Pedigree A schematic representation of the filiation relationship in a family. When the family is small the term micro-pedigree is often used.

Penetrance The fraction of individuals of a given genotype that effectively exhibit the expected phenotype. Penetrance is usually expressed as a percentage. Where less than 100% of genotypically mutant animals are phenotypically mutant, the phenotype is said to be *incompletely penetrant*. The determinism of penetrance is not known. In most cases it results from chance (developmental noise) but can also be influenced by *modifier genes*.

Pericentric In the vicinity of the centromere or involving the centromere – example: a pericentric inversion (see Chap. 3).

PFGE or Pulsed-field gel electrophoresis A technique for separating large DNA molecules from each other (see Chap. 5).

Phenotype The physical manifestation of a genotype within an animal. A mutant phenotype is caused by a mutant genotype and is manifested as an alteration within an animal that distinguishes it from the wild-type. Phenotypes range from severe malformations leading to death or debility to extremely subtle changes in the physical properties of a biological molecule (for example its electrophoretic charge).

Phenotypic marker Phenotypes for which the variation observed in a population is entirely explained by a single "mendelian" factor.

Phylogenetic tree A diagram showing the postulated evolutionary relationships that exist among related species in terms of their divergence from a series of common ancestors at different points in time (see Chap. 1).

Physical map A map based on a great number of minimally overlapping cloned DNAs.

Pleiotropy Pleiotropy describes a situation where a mutant allele has an effect on different (apparently unrelated) phenotypic traits. Mice homozygous for the

piebald allele (*Ednrb*^{s}) have defects in pigmentation, are deaf and often die from megacolon. Piebald has pleiotropic effects.

Poly-A tail A stretch of poly(A) added at the 3' end of mRNAs during transcript maturation and before splicing. The poly-A string ensures the stability of the transcript.

Polygenic A phenotype resulting from the interactions of two or more genes with alternative alleles (see Chap. 2).

Polymorphic A term formulated by population geneticists to describe loci at which there are two or more alleles that are each present at a frequency of at least 1 % in a population of animals. Then, a polymorphism is a genotypic variation within a population.

Polytypic species A species where several subspecies or geographical/morphological races are recognized. *Mus m. domesticus* is typically a polytypic species.

Position effect Corresponds to the variations in the expression of a gene when its molecular environment is changed either after translocation or through transgenesis (see Chap. 8).

Positional cloning See Forward genetics.

Primary RNA The RNA molecule before splicing.

Primers Short oligonucleotides, which anneal to template DNA to prime PCR.

Promoter See TATA box; CAT-box and 5'UTR.

Proximal A relative term meaning closer to the centromere; the opposite of distal.

Pseudogene A DNA sequence that closely resembles a functional gene but is not expressed. Processed *pseudogenes* do not have introns or promoters. They are copied from mRNA and incorporated into the genome. *Unprocessed pseudogenes*, originate from the retrotranscription of messenger RNAs back into the genomic DNA in more or less random locations (see Chap. 5). Pseudogenes are sometimes extremely difficult to differentiate from real genes and some of them even have a function.

***q*-arm** The long arm of a sub-metacentric chromosome (*q* stands for queue)

QTL—quantitative trait locus (plural: Quantitative trait loci—QTLs) are sequences of DNA containing or linked to the genes that determine a quantitative trait.

Quantitative trait A phenotype that can vary in a quantitative manner when measured among different individuals (see Chap. 10). The variation in expression can be due to combinations of genetic and environmental factors, as well as chance.

Radiation hybrids Somatic cell hybrids with a full set of hamster chromosome and fragments of mouse chromosomes, generated by X or γ-irradiation, randomly

inserted into the hamster chromosomes. These interspecific cell hybrids have been very helpful for the (non-meiotic) chromosomal assignment of cloned genes in the mouse.

Recessive allele A recessive allele expresses its characteristic phenotype only when homozygous.

Reciprocal translocations Reciprocal (or balanced) translocations are rearrangements resulting from a reciprocal exchange between the telomeric ends of two non-homologous chromosomes with no change in the total genomic information content. Reciprocal translocations are the most common form of structural rearrangements of the mouse karyotype.

Recombinant congenic strain (RCS) A variation on recombinant inbred strains in which the initial outcross is followed by several generations of backcrossing prior to inbreeding (see Chap. 9).

Recombinant inbred (RI) strain A special type of inbred strain formed from an initial outcross between two inbred strains followed by at least 20 generations of inbreeding (see Chap. 9).

Recombinant The result of a crossing-over in a doubly heterozygous parent such that alleles at two loci flanking the crossing-over that were present on opposite homologs are put together on the same homolog.

Restriction fragment length polymorphism (RFLP) A DNA variation that affects the distance between contiguous restriction sites (most often a nucleotide change that creates or suppresses a site) within or flanking a DNA fragment that hybridizes to a cloned probe (see Chap. 4). RFLPs are detected upon Southern blot hybridization. This polymorphism has been extensively exploited as a genetic polymorphism.

Retrotransposon or retroposon An inserted genomic element that originated from the reverse transcribed mRNA produced from another region of the genome (see Chap. 5).

Reverse genetics A strategy whose aim is to characterize the function of a gene by analyzing the consequences, at the phenotypic level, of alterations occurring spontaneously or engineered at the DNA level (the opposite of forward genetics).

Robertsonian translocation A fusion between the centromeres of two acrocentric chromosomes producing a single metacentric element (see Chap. 3). Robertsonian translocations reduce the number of centromeres but do not alter the number of chromosome arms.

Round spermatid injection (ROSI) The fertilization of super-ovulated oocytes with the nucleus of round spermatids. When the round spermatid is removed from young males, this technique dramatically reduces the time required for the development of fully congenic mouse strains (see Chap. 2).

Satellite DNA A discrete fraction of DNA visible in a cesium chloride density-gradient as a "satellite" to the main DNA band. The term refers to all simple sequence DNA having a centromeric location (see Chap. 5).

Semidominant An allele is said to be semidominant when the phenotype of the heterozygotes is intermediate between the phenotype for the dominant allele and the recessive allele. The $Kit^{W\text{-}f}$ mutant allele is typically semidominant: $Kit^{W\text{-}f/+}$ mice have a fuzzy coat, while $Kit^{W\text{-}f/W\text{-}f}$ and $Kit^{+}/^{+}$ mice are white and fully pigmented respectively.

Shotgun sequencing A method of sequencing DNA that does not require the physical mapping of large sized cloned fragments. For shotgun sequencing, the DNA if fragmented mechanically (i.e. randomly) into segments with a size ranging from 100 to 1,000 bp. The fragments are then sequenced using the chain termination method. Finally the individual sequences are ordered in silico into a continuous sequence based on the sequence overlapping (see Chap. 5).

Silencer A DNA sequence capable of binding regulatory sequences.

Silent (or synonymous) substitution A mere SNP with no effect on the protein sequence. For example, the codons CCT and CCC both code for the proline amino acid.

Simple sequence length polymorphism (SSLP) The polymorphism at a microsatellite locus. Also called SSR for "simple sequence repeats".

SINE Short interspersed element. Families of selfish DNA elements that are a few hundred base pairs in size and dispersed throughout the genome (see Chap. 5).

Single nucleotide polymorphism (SNP) A one base-pair difference between two DNA sequences that is either natural or induced. SNPs can be used as molecular markers.

Spliceosome A set of highly specific molecules (at least five small nuclear RNAs and around 150 proteins) that is essential for RNA splicing (see Chap. 5).

Splicing A mechanism leading to the excision of the introns, i.e. the regions flanked by a splicing donor site and a splicing acceptor site (see Chap. 5).

spretus Abbreviated form of *Mus spretus*, a mouse species commonly used in interspecific matings for the generation of high-density/high-resolution linkage maps (see Chaps. 2 and 9). This species is common in the western Mediterranean border.

SSCP (single strand conformation polymorphism) A sensitive gel-based technique (different from denaturing gradient gel electrophoresis, DGGE) for detecting single nucleotide changes within allelic polymerase chain reaction (PCR) products that have been denatured and gel-fractionated as single strands.

SSLP Simple Sequence Length Polymorphism; SSR: Simple Sequence Repeat; and STR: Short Tandem Repeat.; see microsatellite.

Strain distribution pattern (SDP) The distribution of the segregating alleles at a single locus across a group of animal used for analysis in a linkage study (see Chap. 9). Used in the context of backcross data and data obtained from recombinant inbred (RI) strains.

Strain Refers to a population of mice with known lineage that are bred within a closed colony in order to maintain certain defining characteristics. Inbred strains are produced by brother-sister matings. Random-bred colonies are called outbred stocks (see Chap. 3).

Superovulation Superovulation of female mice allows harvesting large numbers of fertilized eggs or unfertilized oocytes for the purpose of experimental manipulation. It requires the injection of females aged 3 to 5 weeks with gonadotropin hormones that artificially induces ovulation.

Sympatric Refers to related species that have overlapping ranges in nature but do not interbreed. In different parts of its range, *Mus musculus* is sympatric with *Mus macedonicus*, *Mus spicilegus*, and *Mus spretus* (see Chap. 2).

Syngenic Literally "of the same genotype." Used most frequently by immunologists to describe interactions between cells from the same inbred strain.

Syntenic Describes the physical co-localization of genetic loci on the same chromosome within an individual or species. Conserved synteny refers to the situation where two linked loci in one species have homologs that are also linked in another species.

Targeted Mutation A type of mutation in which a DNA construct assembled in vitro substitutes a gene in the genome. The constructs designed to eliminate gene function (loss of function) are often referred to as knockouts (KO), while constructs designed to introduce a mutation in the gene sequence are often referred to as knock-ins (KI).

TATA box The first core promoter sequence identified in eukaryotic protein-coding genes. The TATA box is the binding site of transcription factors or histones. Only 25% of mammalian genes contain a TATA box.

Taxon (plural taxa) Any recognized level in the systematical nomenclature (e.g. species, subspecies…).

Telocentric A chromosome in which the centromere is at one end. Many cytogeneticists consider that telomeric chromosomes, in fact, are acrocentric with a very short arm. True telocentric chromosomes are probably instable structures.

Telomere The distal end of a chromosome. Telomeres consist of repetitive sequences and are considered as insulators whose role is to prevent the chromosome ends of being damaged. Telomeres may play a fundamental role in the control of senescence.

Transgene A fragment of foreign DNA that has been incorporated (randomly) into the genome through the manipulation of pre-implantation embryos. The individuals carrying a transgene are called transgenic (Tg) (see Chap. 6).

Translocation Pertaining to a novel chromosome formed by breakage and reunion of DNA molecules into a new configuration (see Chap. 5).

Transposons Short DNA sequences that have the capacity to change their position within the genome. Some transposons have been engineered to serve as tool for the generation of tagged mutations (see Chap. 7).

Triploïd A conceptus with 3n chromosomes instead of 2n.

Trisomic A conceptus with 2n + 1 chromosomes instead of 2n − symbol Ts. All trisomic mice (except those for chromosome 19) are inviable. Models for human chromosome 21 trisomy have been developed in the mouse (see Chap. 3).

Twins Monozygotic twins are extremely rare in mice, if even they exist.

Ultraconserved Elements (UCE) UCEs are highly conserved DNA sequences shared among evolutionary distant taxa. In most cases, the function(s) of these UCEs is unknown but might not be essential.

Unequal crossover A crossover event that occurs between non-allelic sites. Unequal crossover can lead to the duplication of sequences on one homolog and the deletion of sequences on the other (see Chap. 5). The deleted haplotype is, in most cases, eliminated while the duplicated haplotype generate a CNV.

Variant Literally, an alternative form. Used in conjunction with locus, phenotype, or mouse strain. A 'DNA variant' is equivalent to an alternative DNA allele. A variant mouse usually refers to an animal that carries a mutant allele or expresses a mutant phenotype.

VNTR "Variable number of tandem repeats" locus; see Minisatellite.

Whole-genome shotgun sequencing (WGS) A sequencing strategy that consists of the mechanical fragmentation (e.g., by sonication) of the mammalian DNA into short segments, which are sequenced from both ends using the chain termination method. WGS is fast and less expensive than hierarchical sequencing.

Wild type An allele that functions normally and is commonly found in wild populations.

YAC Yeast artificial chromosome. A vector for cloning genomic inserts from 300 kb to 1 Mb in length. YACs are relatively unreliable vectors, some being deleted and others chimerical. They must then be analyzed with great care.

Zygote The fertilized egg containing pronuclei from both the mother and the father.