Pierre Pontarotti *Editor*

Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life

Springer

Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life

Pierre Pontarotti
Editor

Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life

Springer

Editor
Pierre Pontarotti
CNRS, Laboratoire Évolution Biologique
et Modélisation
Université d'Aix-Marseille
Marseille
France

ISBN 978-3-319-07622-5 ISBN 978-3-319-07623-2 (eBook)
DOI 10.1007/978-3-319-07623-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014943947

Printed on acid-free paper

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

Preface

For the 17th time, the Evolutionary Biology Meeting at Marseilles took place. The goal of this annual meeting is to allow scientists of different disciplines, who share a deep interest in evolutionary biology concepts, knowledge and applications, to meet and exchange and enhance interdisciplinary collaborations.

The Evolutionary Biology Meeting in Marseilles is now recognised internationally as an important exchange platform and a booster for the use of evolutionary-based approaches in biology and also in other scientific areas.

This year more than 100 presentations were selected by the evolutionary biology meeting scientific committee. We have further selected 18 of the most representative ones for the book.

The book will give the reader an overview of the state of the art in the evolutionary biology field. The book is the seventh that we have published further to the meeting. I would like to underline that the seven books are complementary one to another and should be considered as tomes.

The reader of the evolutionary biology books as well as the meeting participants would maybe like me witness years after years during the different meetings and book editions a shift in the evolutionary biology concepts as for example the possibility of HGT in metazoan. The fact that the chapters of the book are selected from a meeting enables the quick diffusion of the novelties.

Concerning the book, the chapters are organised in the following parts:

Molecular and Genome Evolution (Chaps. 1–10)
Phylogeography, Speciation and Coevolution (Chaps. 11–17)
Exobiology and Origin of Life (Chap. 18)

I would like to thank all the authors, the meeting participants, the sponsors: Aix Marseille Université, CNRS, ITMO, ECCOREV FEDERATION, Conseil Général 13, Ville de Marseille.

I also wish to thank Springer's edition staff and in particular Andrea Schlitzberger for her competence and help.

I wish also to thank members of the Association pour l'Etude de l'Evolution Biologique (AEEB) and the members of the Aix Marseille University for their help in the meeting organisation.

Finally, I thank the AEEB manager: Marie Hélène Rome for the organisation of the 17th Evolutionary Biology Meeting and her help with the book.

Marseille, France, April 2014 Pierre Pontarotti

Contents

Part I
Molecular and Genome Evolution

Chapter 1
Comparative Biochemistry and Evolution of Milk Oligosaccharides of Monotremes, Marsupials, and Eutherians

Tadasu Urashima, Michael Messer and Olav T. Oftedal

Abstract The milk saccharides found in monotremes, marsupials, and eutherians show similarities as well as differences. The milk/colostrum of most eutherians, with the notable exception of the Arctoidea (Carnivora), contains lactose as the dominant saccharide plus a variety of oligosaccharides, whereas oligosaccharides predominate over lactose in the milks of monotremes such as the platypus and echidna and marsupials such as the tammar wallaby and koala. In the platypus, some milk oligosaccharides contain Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc (Lacto-*N*-neotetraose, LN*n*T) or Gal(β1-4)GlcNAc(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc (lacto-*N*-neohexaose, LN*n*H) as core structural units as do some milk oligosaccharides of eutherians such as humans, some arctoid carnivores, and elephants. In contrast, milk oligosaccharides of the tammar wallaby and other studied marsupials consist mostly of linear β(1-3)Gal-linked structures ranging in size from tri- to at least decasaccharides, which have never been found in monotreme or eutherian milks, plus GlcNAc-containing oligosaccharides such as Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc (lacto-*N*-novopentaose I). We recently found, however, that Neu5Ac(α2-3)Gal(β1-3)Gal(β1-4)Glc, Neu5Ac(α2-3)Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc, and Gal(β1-3)[Neu5Ac(α2-6)Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc are common to the milks of a marsupial, the red kangaroo, and a eutherian, the Bactrian camel, even though never detected in the milks of monotremes. From these and other observations, we hypothesize that milk oligosaccharides with core units of lactose, LN*n*T, LN*n*H, and possibly lacto-*N*-novopentaose I, including sialylated forms, were present in the common

T. Urashima (✉)
Graduate School of Animal and Food Hygiene, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido, Japan
e-mail: urashima@obihiro.ac.jp

M. Messer
School of Molecular Bioscience, The University of Sydney, Sydney, NSW, Australia

O. T. Oftedal
Smithsonian Environmental Research Center, Edgewater, MD, USA

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_1,

ancestor of extant mammals. If so, the ability to synthesize these oligosaccharides has subsequently been lost in some mammalian lineages. The ability to synthesize long β(1-3)polygalactose sequences in milk oligosaccharides appears to have been acquired by marsupials after their divergence from eutherians. We also discuss the homology and heterogeneity of the non-reducing termini and sialic acid species in milk oligosaccharides and the distribution across mammals of Gal(β1-4)GlcNAc (type I) and Gal(β1-3)GlcNAc (type II) structural units. It appears that milk oligosaccharides have been subject to changing selective pressures as mammary secretions have fulfilled different functions over the course of evolution.

Abbreviations

Glc	Glucose
Gal	Galactose
GlcNAc	*N*-acetylglucosamine
GalNAc	*N*-acetylgalactosamine
Fuc	Fucose
Neu5Ac	*N*-acetylneuraminic acid
Neu5Gc	*N*-glycolylneuraminic acid
UDP	Uridine 5′-diphosphate
CMP	Cytidine 5′-diphosphate

1.1 Introduction

Mammalia is a Class defined by mammary glands and, in reproductive females, the ability to secrete milk. Mammalian milks normally contain major constituents such as water, proteins, lipids, carbohydrates, minerals, and vitamins, but concentrations differ considerably among species. However, unique milk proteins such as caseins, α-lactalbumin, β-lactoglobulin, and whey acidic protein are thought to have evolved via gene replication and modification of ancestral constituents long before the initial appearance of mammals (see Oftedal (2013) for a recent review).

Mammals evolved as a subset of the mammaliaforms (advanced cynodonts) that diversified in the Triassic and Jurassic periods, more than 200 MYBP. Extant mammals consist of monotremes (Prototheria), marsupials (Metatheria), and placentals (Eutheria). The ancestors of the monotremes (australosphenidans) probably diverged from the common theriomorph ancestors of Metatheria and Eutheria during the early Jurassic about 190 MYBP, while the Metatheria and Eutheria diverged somewhat later (ca. 160 MYBP) in the Jurassic (Messer et al. 1998; Luo et al. 2011).

A unique characteristic of monotremes is their oviparous mode of reproduction: their young are hatched from eggs instead of emerging from a uterus. This is

Fig. 1.1 Exposed pouch of lactating Tasmanian echidna showing two milk patches as indicated. The 5c Australian coin depicting an echidna is 2 cm in diameter. *Photo* kindly supplied by Dr. Stewart Nicol, School of Zoology, University of Tasmania

considered a primitive characteristic as egg-laying is a shared feature with other amniotes, such as lizards, turtles, non-avian dinosaurs, and birds. Monotremes have a number of other ancestral characteristics not shared by marsupials and placentals, such as cervical ribs and dwarf nephrons. Extant monotremes comprise only three genera. The platypus, *Ornithorhynchus anatinus*, is found only in eastern Australia, including Tasmania. The second extant monotreme, which is found in both Australia and New Guinea, is the short-beaked echidna *Tachyglossus aculeatus*. The third genus of long-beaked echidnas, *Zaglossus*, includes three species found only in New Guinea, but is not well studied. In monotremes the milk does not emerge from nipples but from two areas of the skin called mammary areolae or milk patches, each of which contains about 100–200 ostia for mammary secretion (Fig. 1.1). Each ostium opens into the infundibulum of a mammary hair follicle.

The term "marsupial" is based on the fact that the young often, but not always, suckle from inside a pouch or marsupium. In marsupials the milk is conveyed via nipples which, depending on species, may or may not be located inside a pouch. The mammary secretory tissue increases greatly in volume during the course of lactation. There are altogether about 330 species of extant marsupials, distributed among seven orders (Wilson and Reeder 2005). Nearly all marsupials are found in either South America or Australasia (Australia and eastern Indonesia and New Guinea); only the Virginia opossum has managed to colonize much of North America. Some diprotodont marsupials, such as kangaroos, can produce two kinds of milk simultaneously, one being secreted from a small teat, supplying a newborn young, while the other is secreted from a large teat, supplying a much older mobile offspring which continues to suckle from outside the pouch.

One of the major differences between marsupials and eutherians lies in their patterns of reproduction. In marsupials, gestation is usually short compared with eutherians of similar size and, as a result, marsupials at birth are very altricial, i.e., small and mostly undeveloped, weighing only between 10 and 750 mg, depending

on the species. Thus, in many developmental features the newborn marsupial (and hatchling monotreme) resembles eutherian embryos; most development and growth occurs after birth during a long period of lactation.

Although the three infraclasses of mammals are very different in their modes of reproduction, they have in common the fact that their young are suckled on milk and that this milk is produced from mammary glands. With very few exceptions the milk of all mammals contains the disaccharide lactose as well as oligosaccharides, which contain a lactose unit at their reducing ends (Urashima et al. 2001a; Messer and Urashima 2002).

This chapter focuses on the chemical structures of oligosaccharides in the milks of various species and elaborates on a hypothesis concerning the evolution of these structures, first enunciated in 2002 (Messer and Urashima 2002).

1.2 Lactose and Milk Oligosaccharides, Function of Milk Oligosaccharides

The milk of most eutherian mammals contains lactose (Gal(β1-4)Glc) as a predominant saccharide (Jenness et al. 1964). When the young consume their mother's milk, the lactose is split into galactose and glucose by intestinal lactase (neutral β-galactosidase or lactose-phlorizin hydrolase), which is located in the membrane of the microvilli of the brush border of the small intestine, and the monosaccharides are transported into the enterocytes by specific pathways. Most of the glucose subsequently enters the circulation and can be used as a source of energy and for synthetic purposes, while most of the galactose is converted to glucose in the liver with the same ultimate fates.

In addition to lactose the milks of monotremes, marsupials, and eutherians also contain trace levels of free monosaccharides plus other larger saccharides, the milk oligosaccharides. Nearly all of these oligosaccharides contain lactose at their reducing ends, to which *N*-acetylglucosamine (GlcNAc), *N*-acetylgalactosamine (GalNAc), galactose (Gal), fucose (Fuc) and/or various sialic acid residues have been attached to form complex structures. It is thought that milk oligosaccharide synthesis is initiated within lactating mammary glands by the actions of several glycosyltransferases that utilize lactose as an acceptor (Messer and Urashima 2002). Human milk contains about 60 g/L of lactose along with 12–13 g/L of more than 200 different milk oligosaccharides (Newburg and Naubauer 1995; Ninonuevo et al. 2006). Thus in the milk of humans and that of many other eutherians lactose predominates over milk oligosaccharides. In contrast, milk oligosaccharides predominate over lactose in the milks of monotremes, marsupials, and some species of the order Carnivora among eutherians (Urashima et al. 2001a; Messer and Urashima 2002; Oftedal 2013).

In eutherian neonates lactose is generally degraded by brush border small intestinal lactase and is utilized as an important energy source. In humans most

(but not all; see below) of the milk oligosaccharides remain intact until they reach the colon because of the presumed absence of other intestinal glycosidases such as fucosidase. Human milk oligosaccharides (HMO) are not degraded or are degraded very slowly by brush border, salivary, and secreted pancreatic enzymes from pigs and adult humans (Engfer et al. 2000; Gnoth et al. 2000), but it is worth remembering that human infant enzymes have not been tested. Therefore, in eutherians milk oligosaccharides appear to have biological significance other than purely nutritional (Urashima et al. 2001a; Messer and Urashima 2002). In monotremes and marsupials the available evidence suggests that, in view of the absence of relevant brush border glycosidases in the small intestine, milk oligosaccharides are transported intact into the small intestinal enterocytes via pinocytosis or endocytosis or perhaps even a paracellular pathway, and are hydrolyzed within intracellular lysosomes by acid glycosidases, after which the resulting monosaccharides are metabolized or enter the circulation. Thus, it appears that in these mammals the digestion/absorption of milk oligosaccharides takes place via a mechanism that is very different from that for lactose in eutherians (Walcott and Messer 1980; Stewart et al. 1983; Crisp et al. 1989a).

The biological functions of milk oligosaccharides have been studied mainly in humans, both in vitro and in vivo. There is evidence that HMO stimulate the growth and colonization of beneficial bifidobacteria in the infant colon, thus acting as prebiotics. Of several bifidobacterial species which colonize the infant colon, *Bifidobactrium longum* subsp. *infantis* and *Bifidobacterium bifidum* have been shown to grow well with HMO as a sole carbon source (Asakuma et al. 2011).

It is assumed that the glycosyltransferases that catalyze the synthesis of HMO also catalyze the biosynthesis of the carbohydrate moieties of glycoproteins and glycolipids in the mammary gland and other tissues. Therefore, it is thought that the structures of the non-reducing ends of HMO are similar to those of the sugar chains of glycoconjugates that are found on the surface of epithelial cells. It is known that infection by many pathogenic bacteria and viruses begins by binding to specific sugar chains of glycoconjugates on the surface of the mucous epithelium of the digestive and respiratory tracts. Therefore, human milk oligosaccharides could be useful for elucidating the structures of the sugar chains that are the targets of these microorganisms. One example is the action of *Campylobacter jejuni*, which can produce serious diarrhea in infants (Ruiz-Palacios et al. 2003).

It was noted above that although the major part of HMO resist digestion and absorption within the infant small intestine, a small proportion of HMO may be absorbed. Unequivocal detection of HMO in the blood of infants has not yet been reported (Bode 2006). Nevertheless, it seems very likely, based on their observed urinary excretion (Rudloff et al. 1996; Obermeier et al. 1999), that small amounts of intact milk oligosaccharides are normally absorbed from the gastrointestinal tract, and that they are transported into the systemic circulation. It follows that they may also alter protein–carbohydrate interactions at a systemic level. Recent studies suggest that HMO interfere with the adhesion of neutrophils to vascular endothelial cells (Klein et al. 2000) and platelets (Bode et al. 2004). These effects

appear to be based on the structural resemblance of some HMO to the glycoprotein ligands of selectins.

In addition, a recent study showed that disialyl lacto-*N*-tetraose (Neu5Ac(α2-3)Gal(β1-3)[Neu5Ac(α2-6)]Gal(β1-4)Glc, DSLNT), a HMO, protected neonatal rats from necrotizing enterocolitis, one of the most common and often fatal intestinal disorders in preterm infants (Jantscher-Krenn et al. 2012).

In the light of the above, it appears that in humans and possibly other eutherians, milk oligosaccharides have significant biological functions other than nutritional ones. These functions could be especially significant for the neonates of bears and other species of Carnivora whose milk is rich in oligosaccharides relative to lactose.

1.3 Biosynthesis of Lactose: The Molecular Evolution of α-Lactalbumin from Lysozyme

Lactose is synthesized within lactating mammary glands from UDP-Gal (donor) and glucose (acceptor) by a transgalactosylation catalyzed by lactose synthase, whereas milk oligosaccharides are synthesized by several specific glycosyltransferases that transfer monosaccharide residues to lactose and/or small lactose-based oligosaccharides as acceptors. Lactose synthase is a complex of β4galactosyltransferase I (β4GalTI) and α-lactalbumin, which is one of the whey proteins. Tissues and fluids other than lactating mammary glands and milk do not contain α-lactalbumin but do contain a β4GalTI which transfers galactose from UDP-Gal to nonreducing GlcNAc residues in glycoconjugates to synthesize *N*-acetyllactosamine (Gal(β1-4)GlcNAc) units. In lactating mammary glands, the binding of α-lactalbumin to β4GalTI changes its preferred acceptor from GlcNAc to glucose. Thus, the expression of α-lactalbumin within the mammary gland is the key to the presence of lactose in milk.

It has been suggested that the ratio of milk oligosaccharides to lactose is determined mainly by the rate of lactose synthesis (Messer and Urashima 2002). When this rate is low, relative to the activities of the pertinent glycosyltransferases, most of the lactose would be utilized as an acceptor for the glycosyltransferases that synthesize milk oligosaccharides; in consequence the ratio of milk oligosaccharides to lactose would be relatively high. When, however, the rate of lactose production is rapid, only a small proportion of the nascent lactose would function as an acceptor, and thus the ratio of oligosaccharides to lactose would be relatively low. As indicated above, the rate at which lactose is synthesized is primarily controlled by the expression level of α-lactalbumin within the mammary gland, although upregulation of β4GalTI by a mammary specific mechanism is also important (Shaper et al. 1998).

It follows that in most eutherian species, in whose milk lactose is the predominant saccharide, the level of expression of α-lactalbumin within the mammary

glands is likely to be higher, relative to expression and activity of glycosyltransferases, than in monotremes, marsupials and those species of Carnivora in which milk oligosaccharides predominate over lactose (Messer and Urashima 2002).

α-Lactalbumin has been found only in milk or lactating mammary glands. Lysozymes, which cleave the bond in peptidoglycans of bacterial cell walls, are found in many tissues of other vertebrates and in insects. Since the discovery that the amino acid sequence of α-lactalbumin is similar to that of c type lysozyme, this similarity was found to extend to their secondary and tertiary structures and the arrangement of their genes (McKenzie and White 1991). As α-lactalbumin is found only in mammals it must have evolved from c-lysozyme, which has an older phylogenetic origin, not vice versa. It is generally believed that this evolution was initiated by duplication of the c-lysozyme gene. The hypothesis that the high ratio of milk oligosaccharides to lactose in monotremes is due to low expression of α-lactalbumin within the mammary gland is supported by the finding that α-lactalbumin concentration is low in the milks of platypus and echidna, when compared with other milks (Messer et al. 1997; Shaw et al. 1993). It appears likely that the expression of α-lactalbumin would have been low within the mammary glands of the common ancestor of mammals, and that milk oligosaccharides predominated over lactose, as in extant monotremes (Messer and Urashima 2002). If so, subsequent evolutionary increase in α-lactalbumin expression in mammary glands must have occurred in the eutherian lineage since most extant eutherian species secrete lactose as the predominant milk saccharide (Messer and Urashima 2002).

In extant mammals, the lactase activity in the brush border of the small intestine is found only in eutherians and not in macropods (Messer et al. 1989), or other examined marsupials (except for the late lactation stage of the brush-tailed possum) (Crisp et al. 1989a), nor in the short-beaked echidna, a monotreme (Stewart et al. 1983). This suggests that small intestinal brush border lactase activity was acquired during the later divergence of mammalian taxa, as it would have been required before the lactose content of the milk could increase markedly (Oftedal 2013).

One can speculate that oligosaccharides acted mainly as anti-infection components in the milk of the presumably oviparous ancestor of mammals, thus providing protection to the hatchlings (Oftedal 2013; Messer and Urashima 2002). The evolution of α-lactalbumin from lysozyme would have had a selective advantage in that it permitted the synthesis of lactose-based anti-pathogenic oligosaccharides that supplemented the anti-infection function of the original secretory constituent, c-lysozyme.

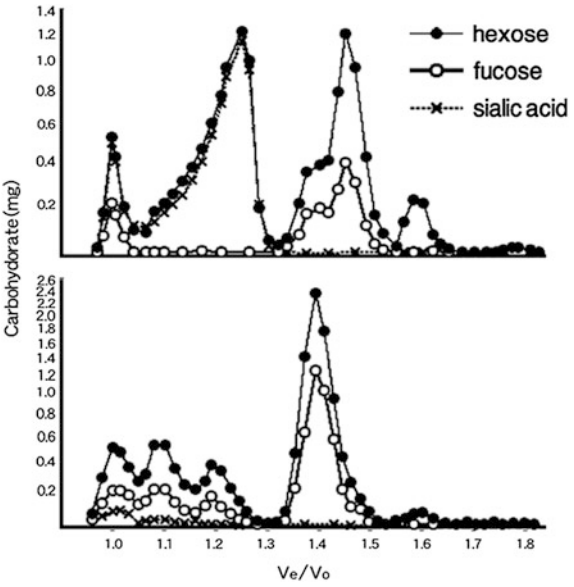


Fig. 1.2 Gel chromatograms of the carbohydrate fractions from milk of echidna and platypus on Sephadex G-15. Fractions were analyzed for hexose, fucose, and sialic acid; *Ve* is the elution volume and V_0 is the void volume. These *figures* are reproduced from Messer and Kerry (1973) with permission

1.4 Structures of Oligosaccharides in Extant Mammals

1.4.1 Oligosaccharides in Monotreme Milk

In 1973, Messer and Kerry separated the milk sugars of the short-beaked echidna, whose milk had been collected on Kangaroo Island and New South Wales, and of the platypus, by chromatography on Sephadex G-15 (see Fig. 1.2) (Messer and Kerry 1973). The four peaks of the echidna carbohydrate fraction corresponded to lactose, fucosyllactose, difucosyllactose, and sialyllactose. The elution profile of platypus milk sugars (Messer et al. 1983) confirmed that they consisted of difucosyllactose and a number of larger saccharides, all of which contained fucose, and very little free lactose. Thus, in the milk of both monotremes, oligosaccharides, and not free lactose were found to be the major carbohydrates.

Subsequent studies on the structure of echidna sialyllactose showed that it uniquely contains an *O*-acetyl group at carbon 4 of the *N*-acetylneuraminic acid (Messer 1974; Kamerling et al. 1982), while the detailed structures of fucosyllactose and difucosyllactose of both monotremes were examined using ^{13}C-NMR (carbon 13 nuclear magnetic resonance spectroscopy) (Jenkins et al. 1984). The structures of the neutral oligosaccharides of platypus milk were further studied in detail by Amano et al. (1985) who used sequential exoglycosidase digestion and methylation. The results of all studies are shown in Fig. 1.3. Platypus acidic oligosaccharides have not so far been characterized. It is noteworthy that some platypus oligosaccharides contain Lewis x (Gal(β1-4)[Fuc(α1-3)]GlcNAc) or Lewis y (Fuc(α1-2)Gal(β1-4)[Fuc(α1-3)]GlcNAc) units.

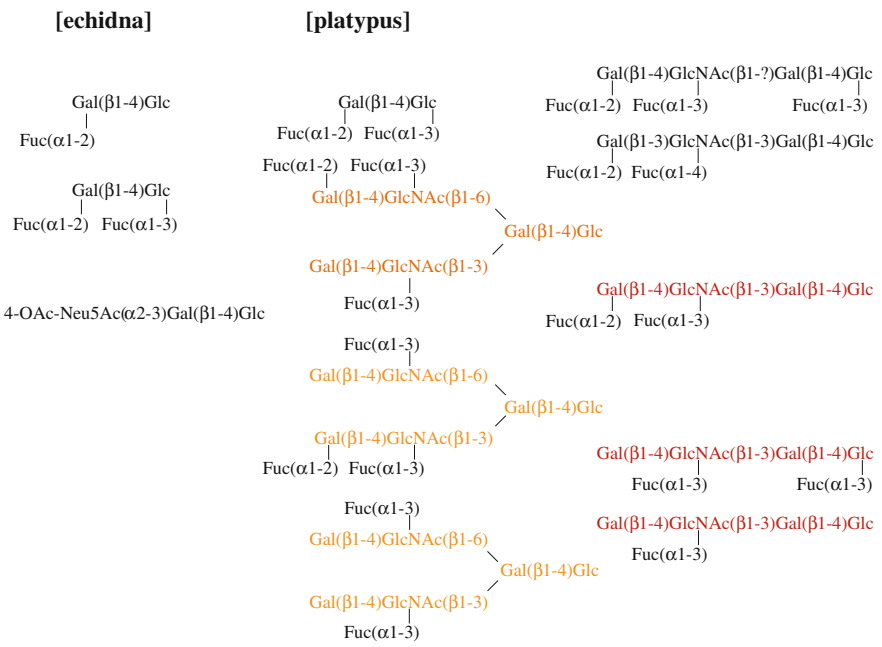


Fig. 1.3 Structures of echidna and platypus milk oligosaccharides. The core units of lacto-*N*-neotetraose and lacto-*N*-neohexaose are shown in *red* and *orange*, respectively (Messer 1974; Kamerling et al. 1982; Jenkins et al. 1984; Amano et al. 1985)

1.4.2 Oligosaccharides in Marsupial Milk

Until recently the tammar wallaby was the only marsupial species whose milk oligosaccharides had been characterized in detail. Figure 1.4 shows a gel chromatogram of tammar wallaby milk carbohydrate on BioGel P-2 (see also Messer and Green 1979). The peak marked with * corresponds to lactose, illustrating that milk oligosaccharides predominate over lactose as in monotremes. The chemical structures of the tammar wallaby neutral oligosaccharides were characterized by ^{13}C-NMR (Messer et al. 1980, 1982; Collins et al. 1981; Bradbury et al. 1983) as in Fig. 1.5. They can be classified into two series. The first major series, whose members are unbranched, can be expressed as [Gal(β1-3)]nGal(β1-4)Glc. Those of the second, minor, series are branched saccharides containing *N*-acetylglucosamine. Subsequently, red kangaroo acidic oligosaccharides and koala oligosaccharides have been characterized (Anraku et al. 2012; Urashima et al. 2013), as described below.

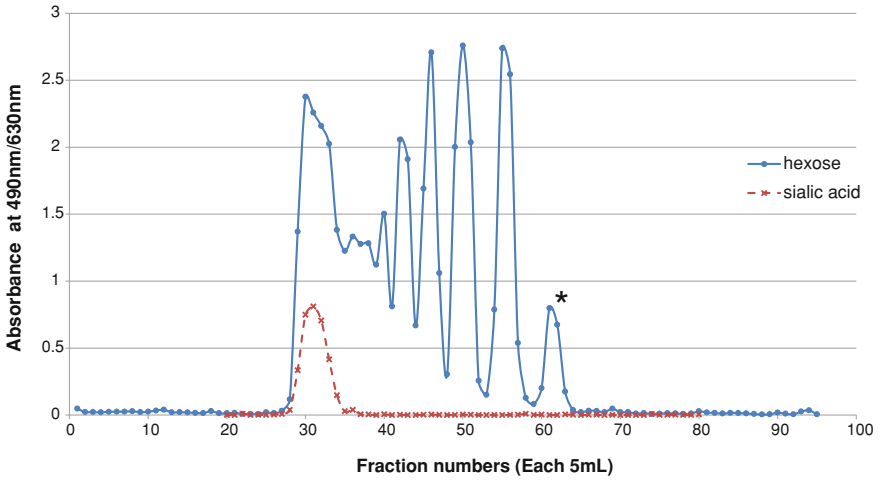


Fig. 1.4 Gel chromatogram of the carbohydrate fraction from tammar wallaby milk on BioGel P-2. The *peak* marked with * corresponds to lactose. Each fraction was monitored by the phenol-H_2SO_4 method at 490 nm for hexose and the periodate-resorcinol method at 630 nm for sialic acid

Fig. 1.5 Structures of tammar wallaby neutral milk oligosaccharides. The core unit of lacto-*N*-novopentaose I is shown in *blue* (Messer et al. 1980, 1982; Collins et al. 1981; Bradbury et al. 1983)

① major series oligosaccharides

Gal(β1-3)Gal(β1-4)Glc

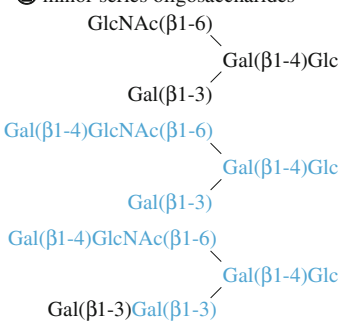
Gal(β1-3)Gal(β1-3)Gal(β1-4)Glc

Gal(β1-3)Gal(β1-3)Gal(β1-3)Gal(β1-4)Glc

Gal(β1-3)Gal(β1-3)Gal(β1-3)Gal(β1-3)Gal(β1-4)Glc

Gal(β1-3)Gal(β1-3)Gal(β1-3)Gal(β1-3)Gal(β1-3)Gal(β1-4)Glc

② minor series oligosaccharides



1.4.3 Oligosaccharides in the Milk and Colostrum of Eutherian Mammals

Among eutherians, oligosaccharides have been characterized in the milks of primates (Urashima et al. 1999b, 2009; Goto et al. 2010; Taufik et al. 2012; Tao et al. 2011) including humans (Amano et al. 2009), artiodactyls (Saito et al. 1984, 1987; Urashima et al. 1989a, 1991a, 1994, 1997a; Mariño et al. 2011; Nakamura et al. 1998; Fukuda et al. 2010; Alhaj et al. 2013; Taufik et al. 2014), perissodactyls (Urashima et al. 1989b, 1991b; Nakamura et al. 2001), carnivorans (Bubb et al. 1999: Urashima et al. 1997b, 1999a, c, 2000, 2001b, 2003a, b, 2004a, b, 2005; Nakamura et al. 2003; Taufik et al. 2013; Kinoshita et al. 2009; Uemura et al. 2009; Senda et al. 2010), proboscideans (Uemura et al. 2006; Osthoff et al. 2008), rodents (Sturman et al. 1985; Choi and Carubelli 1968; Barra and Caputto 1965), cetaceans (Urashima et al. 2002; Uemura et al. 2005; Urashima et al. 2007), bats (Senda et al. 2011), and giant anteaters (Urashima et al. 2008). Here, we will discuss the milk oligosaccharides of species with great oligosaccharide diversity including humans, Japanese black bear (Carnivora), and African elephant (Proboscidea).

Human mature milk and colostrum contain 12–13 and 22–24 g/L of oligosaccharides, respectively, and at least 117 varieties of HMO have been characterized to date. The HMO are classified into 13 groups based on core structures as shown in Fig. 1.6 (Urashima et al. 2011, 2012). The Fuc can attach to Gal via an α(1-2) linkage or to GlcNAc or Glc via α(1-3)/(1-4) linkage of either core structure, while Neu5Ac can attach to Gal via α(2-3)/(2-6) linkage or to GlcNAc via α(2-6) linkage. The oligosaccharide patterns in milk differ depending on the Lewis blood group of the donor. Le (a− , b+) donor (secretor) milk contains all HMO, while Le (a+ , b−) donor (nonsecretor) milks do not contain oligosaccharides which have a nonreducing α(1-2) linked Fuc residue, such as Fuc(α1-2)Gal(β1-4)Glc (2'-FL) and Fuc(α1-2)Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)Glc (LNFP-I). Le (a− . b−) donor (Lewis negative) milks do not contain oligosaccharides which have an α(1-4) linked Fuc residue, such as Gal(β1-3)[Fuc(α1-4)]GlcNAc(β1-3)Gal(β1-4)Glc (LNFP-II). The percentages of secretor, nonsecretor, and Lewis negative donors are estimated to be 80, 15, and 5 %, respectively.

The structures of milk oligosaccharides of Japanese black bear (Urashima et al. 1999c, 2004b) and African elephant (Osthoff et al. 2008) are shown in Figs. 1.7 and 1.8. A significant feature of some Japanese black bear milk oligosaccharides is that they contain B antigen (Gal(α1-3)[Fuc(α1-2)]Gal), α-Gal epitope (Gal(α1-3)Gal(β1-4)Glc(NAc)), and/or Lewis x antigen. Some African elephant milk oligosaccharides contain α-Gal epitope, Lewis x or sialyl Lewis x epitope (Neu5Ac(α2-3)Gal(β1-4)[Fuc(α1-3)]GlcNAc).

Gal(β1-4)Glc	Lactose
Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)Glc	Lacto-*N*-tetraose
Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc	Lacto-*N*-neotetraose
Gal(β1-4)GlcNAc(β1-6) Gal(β1-3)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	Lacto-*N*-hexaose
Gal(β1-4)GlcNAc(β1-6) Gal(β1-4)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	Lacto-*N*-neohexaose
Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc	*para*-Lacto-*N*-hexaose
Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc	*para*-Lacto-*N*-neohexaose
Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-6) Gal(β1-3)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	Lacto-*N*-octaose
Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-6) Gal(β1-4)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	Lacto-*N*-neooctaose
Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-6) Gal(β1-3)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	*iso*-Lacto-*N*-octaose
Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc	*para*-Lacto-*N*-octaose
Gal(β1-4)GlcNAc(β1-6) Gal(β1-3)GlcNAc(β1-3) ⟩Gal(β1-4)GlcNAc(β1-6) Gal(β1-3)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	Lacto-*N*-decaose
Gal(β1-4)GlcNAc(β1-6) Gal(β1-4)GlcNAc(β1-3) ⟩Gal(β1-4)GlcNAc(β1-6) Gal(β1-3)GlcNAc(β1-3) ⟩Gal(β1-4)Glc	Lacto-*N*-neodecaose

Fig. 1.6 The 13 core structures of human milk oligosaccharides

1.4.4 Comparison of the Core Structures of the Milk Oligosaccharides of Monotremes, Marsupials, and Eutherians

Comparisons of the milk oligosaccharides of the platypus and of humans revealed common core structures of milk oligosaccharides for lacto-*N*-neotetraose (Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc, LN*n*T) and lacto-*N*-neohexaose (Gal(β1-4)GlcNAc(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc, LN*n*H) (Amano et al. 1985; Urashima et al. 2011, 2012). On the other hand, these structures were not found in tammar wallaby milk which contains other oligosaccharides such as linear β(1-3) linked galactooligosaccharides and lacto-*N*-novopentaose I (Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc) (Messer et al. 1980, 1982; Collins et al. 1981; Bradbury et al. 1983). Oligosaccharides larger than the trisaccharide of the β(1-3) galactooligosaccharide series have not been found in the milk/colostrum of eutherians or monotremes. Although fucosyl milk oligosaccharides are a significant feature of the milk of monotremes and some eutherians, these had not until recently been found in any marsupial milk. Messer and Urashima (2002) therefore noted that in this respect eutherian milk oligosaccharides are more similar to those of monotremes than to those of marsupials. However, recent studies on the milk oligosaccharides of red kangaroo (diprotodont marsupial) (Anraku et al. 2012) and Bactrian camel (artiodactyl eutherian) (Fukuda et al. 2010) show that marsupial and eutherian milk oligosaccharides share at least one common oligosaccharide.

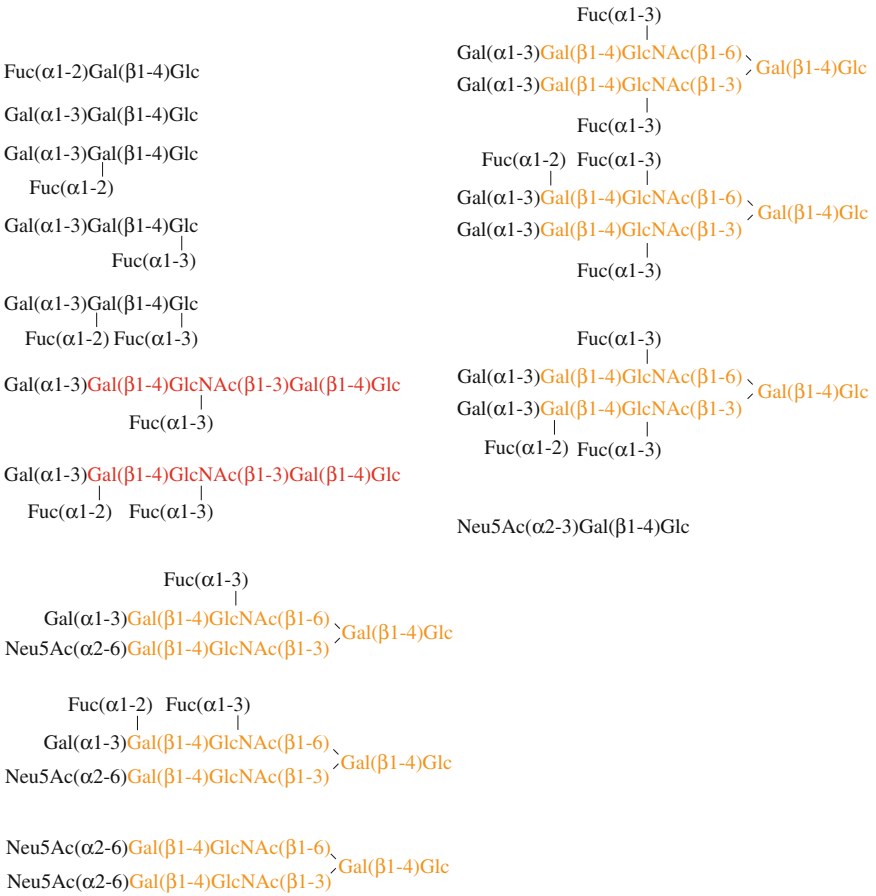


Fig. 1.7 Structures of Japanese black bear milk oligosaccharides (Urashima et al. 1999c, 2004b)

1.4.5 Comparison of Milk Oligosaccharides Between Red Kangaroo and Bactrian Camel

The carbohydrate fractions from red kangaroo milk (Anraku et al. 2012) and Bactrian camel colostrum (Fukuda et al. 2010) were separated by extraction with chloroform/methanol (2:1, v/v), and each oligosaccharide was then purified by gel filtration and anion exchange chromatography followed by normal phase high-performance liquid chromatography (HPLC) with an Amide-80 column. The oligosaccharide structures were characterized by ^{1}H-NMR (proton nuclear magnetic resonance spectroscopy) and MALDI-TOFMS (matrix assisted laser desorption/ionization time of flight mass spectrometry). The HPLC of the acidic oligosaccharide fractions of red kangaroo and Bactrian camel are shown in Figs. 1.9 and 1.10, respectively.

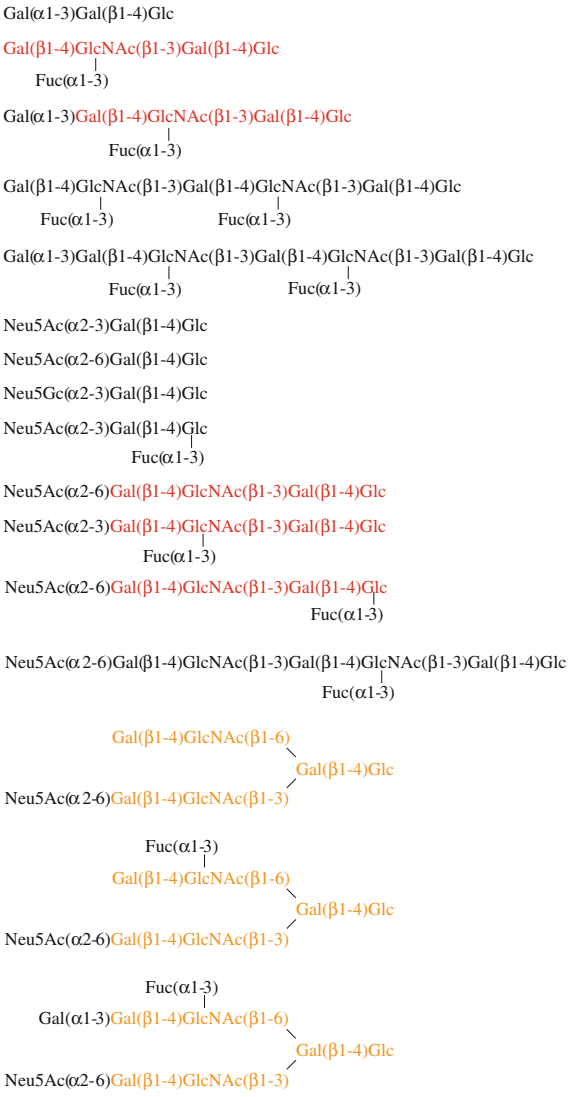


Fig. 1.8 Structures of African elephant milk oligosaccharides (Osthoff et al. 2008)

The chemical structures of red kangaroo milk oligosaccharides are shown in Fig. 1.11. In some oligosaccharides the nonreducing Gal of the linear oligosaccharides, expressed as [Gal(β1-3)]nGal(β1-4)Glc, is substituted by a Neu5Ac residues via α(2-3) linkage or substituted by sulfate at OH-3. In other oligosaccharides, α(2-3) Neu5Ac or sulfate are linked to nonreducing Gal of lacto-*N*-novopentaose I. It is evident that the core structures of some of the red kangaroo acidic oligosaccharides are the same as those of tammar wallaby neutral milk oligosaccharides.

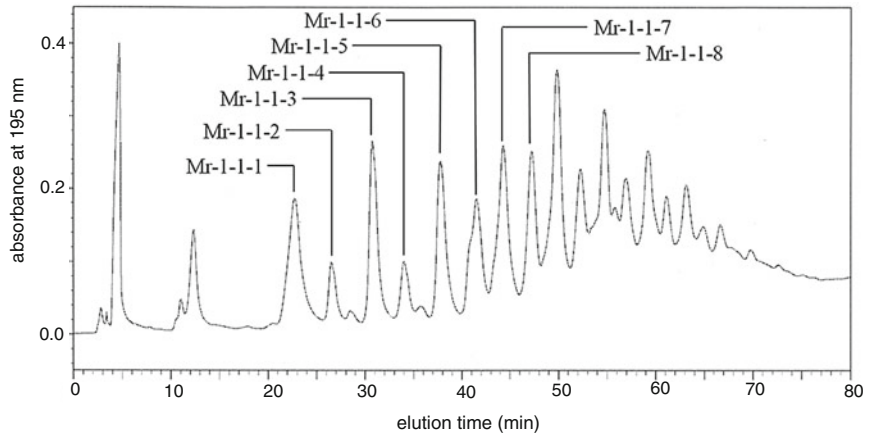


Fig. 1.9 Normal phase high-performance liquid chromatography of the acidic oligosaccharide fraction of red kangaroo milk with Amide-80 column. This figure is reproduced from Anraku et al. (2012) with permission

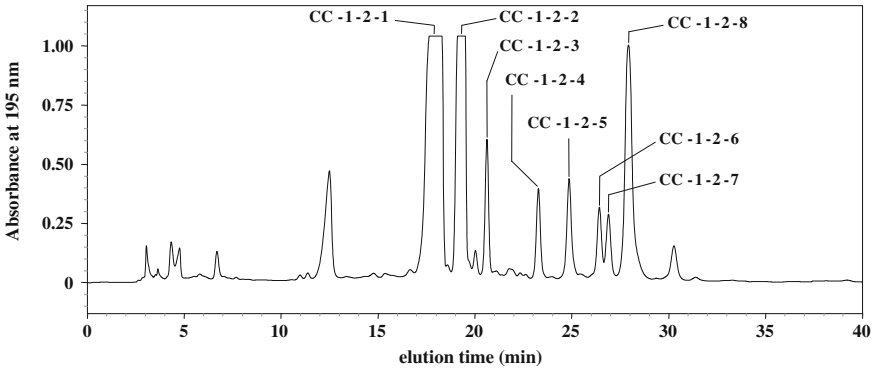


Fig. 1.10 Normal phase high-performance liquid chromatography of the acidic milk oligosaccharide fraction of Bactrian camel colostrum with Amide-80 column. This figure is reproduced from Fukuda et al. (2010) with permission

The neutral and acidic oligosaccharide structures of Bactrian camel colostrum are shown in Fig. 1.12. Sialyl lacto-*N*-neotetraose (Neu5Ac(α2-6)Gal(β1-4)GlcNAc(β1-3)Gal(β1-4)Glc, LST c) and monosialyl lacto-*N*-neohexaose (Neu5Ac(α2-6)Gal(β1-4)GlcNAc(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc), whose core structures are LN*n*T and LN*n*H, respectively, have been also found in milk/colostrum of some eutherians including humans. A recent study showed that the structures of the neutral and acidic oligosaccharides of dromedary camel milk oligosaccharides are essentially similar to those of Bactrian camel oligosaccharides (Alhaj et al. 2013). In summary, it appears that the milk oligosaccharides that are common to both red kangaroo and Bactrian camel are Neu5Ac(α2-3)Gal(β1-3)Gal(β1-4)Glc,

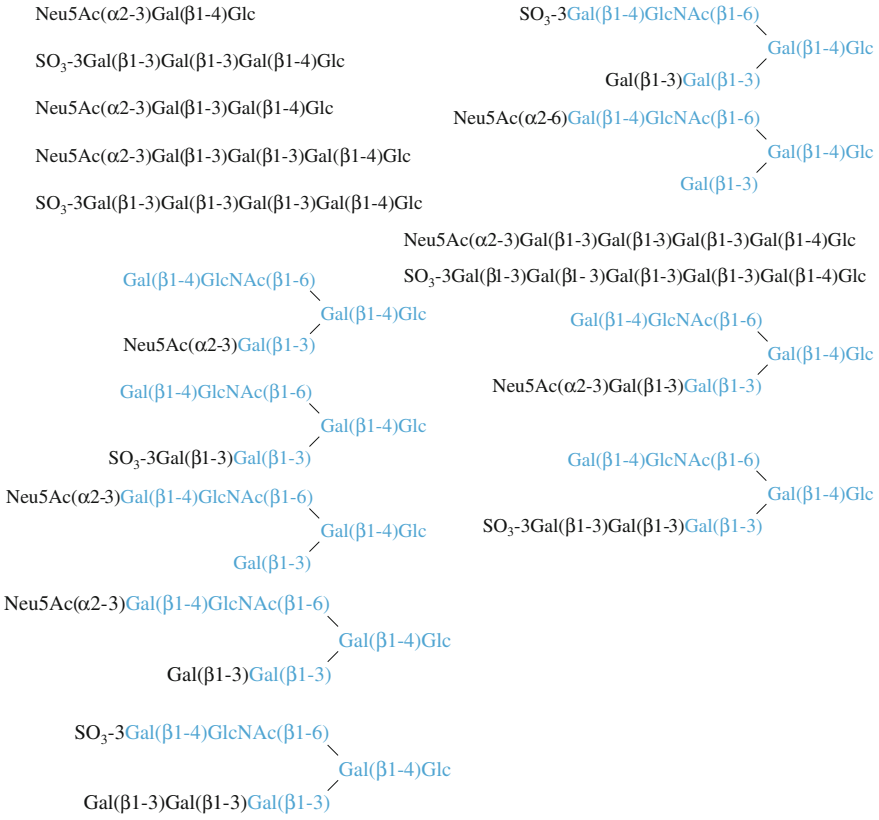


Fig. 1.11 Structures of red kangaroo acidic milk oligosaccharides (Anraku et al. 2012; Urashima et al. 2013)

Neu5Ac(α2-3)Gal(β1-3)[Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc (sialyl lacto-*N*-novopentaose a) and Gal(β1-3)[Neu5Ac(α2-6)Gal(β1-4)GlcNAc(β1-6)]Gal(β1-4)Glc (sialyl lacto-*N*-novopentaose b), whose core structures are Gal(β1-3)Gal(β1-4)Glc or lacto-*N*-novopentaose I (Fig. 1.13).

1.5 A Modified Hypothesis of Milk Oligosaccharide Evolution and a Discussion of the Heterogeneity of Core Structures Among Mammalian Species

Some of the studies have shown that lacto-*N*-novopentaose I and its sialyl derivatives are common to both marsupials and some eutherian milk/colostrum. Lacto-*N*-novopentaose I has been found in milk/colostrum of cow (Urashima et al. 1991a;

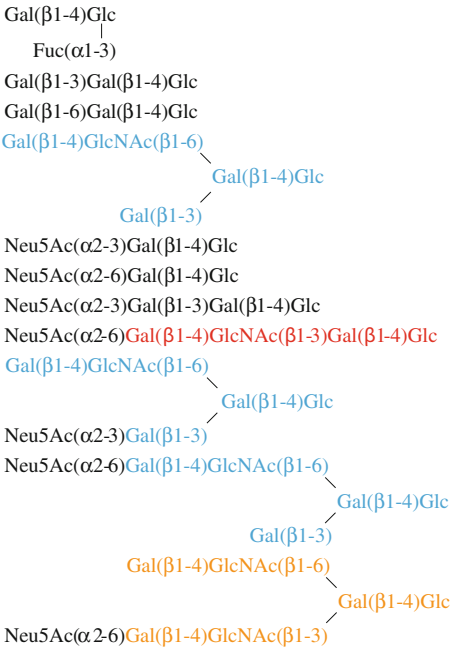


Fig. 1.12 Structures of Bactrian camel milk oligosaccharides (Fukuda et al. 2010)

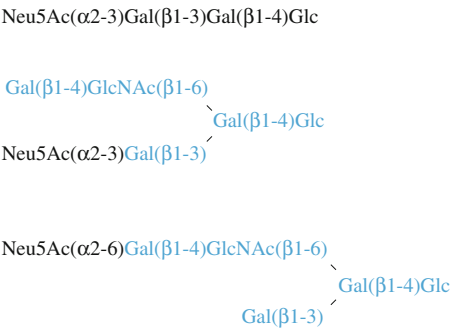


Fig. 1.13 Common milk oligosaccharides between red kangaroo and Bactrian camel

Mariño et al. 2011), horse (Urashima et al. 1989b), pig (Tao et al. 2010), and capuchin monkey (Urashima et al. 1999b; Goto et al. 2010) in addition to Bactrian camel (Fukuda et al. 2010). In addition, Neu5Ac(α2-3){Gal(β1-4)[Fuc(α1-3)] GlcNAc(β1-6)}Gal(β1-4)Glc, whose core is lacto-*N*-novopentaose I, has been found in trace amounts in human milk (Grönberg et al. 1992). This fucosyl oligosaccharide has been recently found in koala milk as well (Urashima et al. 2013).

Based on the above findings we propose the following. Milk oligosaccharides with core structures of LN*n*T, LN*n*H or lacto-*N*-novopentaose I were present in the mammary secretions of the common ancestor of extant monotremes, marsupials,

and eutherians, but oligosaccharides containing lacto-*N*-novopentaose I as a core structure may have been lost during the evolution of monotremes and some eutherian species. Conversely, milk oligosaccharides whose cores were LN*n*T or LN*n*H, may have been lost during the evolution of marsupials. The linear β(1-3) galactooligosaccharides larger than Gal(β1-3)Gal(β1-4)Glc were acquired in marsupials after their divergence from eutherians. Although Messer and Urashima (2002) had noted the absence of fucosyl oligosaccharides from marsupial milks, two such fucosyl oligosaccharides viz Gal(β1-3){Gal(β1-4)[Fuc(α1-3)] GlcNAc(β1-6)}Gal(β1-4)Glc, and Neu5Ac(α2-3)Gal(β1-3){Gal(β1-4)[Fuc(α1-3)] GlcNAc(β1-6)}Gal(β1-4)Glc were recently detected in koala milk (Urashima et al. 2013). It is possible, therefore, that fucosyl oligosaccharides were lost during the evolution of macropods and some other marsupial species but not in the koala. Unfortunately, information on milks secreted by the diverse and speciose groups of marsupials is still very limited. Very little is known about the detailed structures of milk oligosaccharides of taxa normally considered basal, such as the opossums (Didelphimorphia), although Crisp et al. (1989b) demonstrated, using thin layer chromatography (TLC), that the major oligosaccharides in the milk of a South American short-tailed opossum, *Monodelphis domestica*, seemed to be very similar, if not identical, to those found in the tammar wallaby milk. Furthermore, changes in their concentrations during the course of lactation were also similar (Crisp et al. 1989b).

The heterogeneity of core structures of milk oligosaccharides among mammalian species is most likely based on the presence/absence and/or the activity or specificity of some *N*-acetylglucosaminyltransferases within the lactating mammary glands. The biosynthetic pathways of LN*n*T, LN*n*H, and lacto-*N*-novopentaose I are shown in Fig. 1.14. The biosynthesis of LN*n*T begins with the action of iGnT, which transfers GlcNAc from UDP-GlcNAc to the Gal residue of lactose, while that of LN*n*H is due to the action of lGnT, which transfers GlcNAc to the Gal residue of GlcNAc(β1-3)Gal(β1-4)Glc (Kobata 2010). When there is no activity of iGnT within the mammary glands, neither LN*n*T nor LN*n*H can be produced. Therefore, it is likely that the apparent absence of these core units in marsupial milks studied to date is due to absence of iGnT activity in marsupial mammary glands.

Lacto-*N*-novopentaose I is synthesized from Gal(β1-3)[GlcNAc(β1-6)]Gal (β1-4)Glc (lacto-*N*-novotetraose) as a precursor. This tetraose is produced from Gal(β1-3)Gal(β1-4)Glc by β6 *N*-acetylglucosaminyltransferase, which transfers GlcNAc from UDP-GlcNAc to OH-6 of the penultimate Gal of the trisaccharide. Urashima et al. (1992) detected this transferase activity in homogenates of lactating mammary glands of the tammar wallaby. Thus the absence of lacto-*N*-novopentaose I in the milk of monotremes and some eutherian species could be due either to the absence of β6 *N*-acetylglucosaminyltransferase activity in their mammary glands or to the absence of Gal(β1-3)Gal(β1-4)Glc as an acceptor.

In addition to β4galactosyltransferase I activity, Messer and Nicholas (1991) detected β3galactosyltransferase activity, which transfers Gal from UDP-Gal to the nonreducing Gal residue of lactose and of Gal(β1-3)Gal(β1-4)Glc in lactating

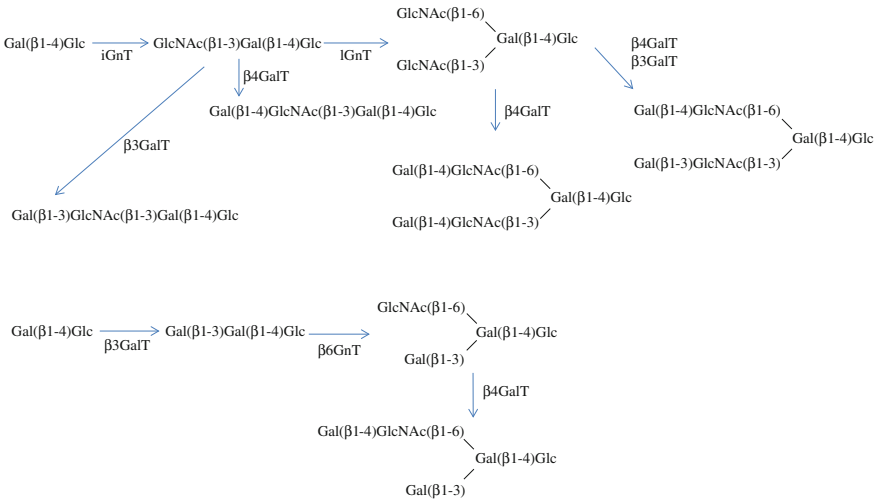


Fig. 1.14 Biosynthetic pathway of lacto-*N*-neotetraose, lacto-*N*-tetraose, lacto-*N*-neohexaose, lacto-*N*-hexaose, and lacto-*N*-novopentaose I

mammary glands of the tammar wallaby. It is likely that the high concentration of the linear β(1-3)galactooligosaccharides and their sialyl derivatives in milk of tammar wallaby or red kangaroo is caused by high expression and activity of this β3 galactosyltransferase in the mammary glands. On the other hand, no β(1-3) galactooligosaccharides larger than 3′-galactosyllactose (Saito et al. 1987; Urashima et al. 1989a, b, 1994; Fukuda et al. 2010; Alhaj et al. 2013; Taufik et al. 2014; Donald and Feeney 1988) and Neu5Ac(α2-3)Gal(β1-3)Gal(β1-4)Glc (Fukuda et al. 2010; Alhaj et al. 2013) have been found in the milk of eutherians. The absence of larger β(1-3)galactooligosaccharides in eutherian milk may be due either to low activity of β3galactosyltransferase in eutherian mammary glands or to a difference in the specificity of this enzyme. We hypothesize that its expression and total activity significantly increased in marsupial mammary glands after the divergence of marsupials from eutherians, thus permitting a greatly increased concentration of carbohydrate in the milk while avoiding a deleterious increase in osmolality (Messer and Green 1979).

1.6 Heterogeneity of Nonreducing Units of Milk Oligosaccharides Among Mammalian Species

Among mammalian species heterogeneity is observed at the nonreducing termini of milk oligosaccharides, as well as in their core structures. Milk oligosaccharides sometimes contain A (GalNAc(α1-3)[Fuc(α1-2)]Gal), B (Gal(α1-3)[Fuc(α1-2)]Gal) and H (Fuc(α1-2)Gal) antigens, α-Gal epitope (Gal(α1-3)Gal(β1-4)Glc(NAc)),

Lewis x (Gal(β1-4)[Fuc(α1-3)]GlcNAc) or Lewis a (Gal(β1-3)[Fuc(α1-4)]GlcNAc) as the terminal nonreducing units.

The presence/absence of these nonreducing units is heterogeneous even among closely related species. For example, in bears (family Ursidae), it was observed that some milk oligosaccharides of Ezo brown bear (Urashima et al. 1997b) and Japanese black bear (Urashima et al. 1999c, 2004b) contain H and B antigens, respectively, while in polar bears some of the milk oligosaccharides of one animal contain A antigen, and of another animal A or B antigens (Urashima et al. 2000). Among carnivoran species other than Ursidae, the presence/absence of ABH antigens in milk oligosaccharides is also heterogeneous depending on species. Milk oligosaccharides containing A antigen including A-tetrasaccharide (GalNAc(α1-3)[Fuc(α1-2)]Gal(β1-4)Glc) have been found in striped skunk (Mephitidae) (Taufik et al. 2013), African lion (Felidae) (Senda et al. 2010) and clouded leopard (Felidae) (Senda et al. 2010), while those containing B antigen including B-tetrasaccharide (Gal(α1-3)[Fuc(α1-2)]Gal(β1-4)Glc) have been found in spotted hyena (Hyaenidae) (Uemura et al. 2009). Those containing H antigen have been found in dog (Canidae) (Bubb et al. 1999), mink (Mustelidae) (Urashima et al. 2005), whitenosed coati (Procyonidae) (Urashima et al. 1999a), hooded seal (Phocidae) (Urashima et al. 2001b), bearded seal (Phocidae) (Urashima et al. 2004a), and harbor seal (Phocidae) (Urashima et al. 2003a). As H antigen is a precursor in the biosynthesis of A and B antigens, milk of species, whose oligosaccharides contain A or B antigen, also contain small amounts of the oligosaccharides containing H antigen including Fuc(α1-2)Gal(β1-4)Glc (2'-fucosyllactose, 2'-FL) (Uemura et al. 2009; Senda et al. 2010; Taufik et al. 2013).

Although oligosaccharides containing the α-Gal epitope including Gal(α1-3)Gal(β1-4)Glc (isoglobotriose) have been detected in the milk of many species of Carnivora (Urashima et al. 1997b, 1999a, c, 2000, 2004b; Nakamura et al. 2003; Taufik et al. 2013; Senda et al. 2010; Uemura et al. 2009), these have not been found in milk of phocid species (Urashima et al. 2001b, 2003a, 2004a) or the dog (Bubb et al. 1999). We hypothesize that the milk of the common ancestor of extant carnivorans contained oligosaccharides with the α-Gal epitope, but that the activity of mammary gland α3galactosyltransferase, which transfers Gal to Gal(β1-4)GlcNAc or lactose, was lost during the evolution of phocids. It is notable that among carnivorans only bears (Ursidae) have Lewis x milk oligosaccharides (Urashima et al. 1997b, 1999c, 2000, 2003b, 2004b).

Studies on the nonreducing terminal units of the milk oligosaccharides of primates have produced some interesting results. A-tetrasaccharide has been found in milk of the bonobo (great ape), while B-tetrasaccharide has been detected in milk of the gorilla (great ape) (Urashima et al. 2009). 2′-FL and/or other oligosaccharides containing H antigen have been found in milk of humans (Urashima et al. 2011, 2012), chimpanzee (great ape), bonobo, gorilla (Urashima et al. 2009), and aye-aye (strepsirrhine primate) (Taufik et al. 2012). Gal(β1-4)[Fuc(α1-3)]Glc (3-fucosyllactose, 3-FL), and/or other oligosaccharides containing Lewis x have been detected in milk/colostrum of humans (Urashima et al. 2011, 2012), chimpanzee, bonobo, orangutan (great ape) (Urashima et al. 2009), rhesus macaque

(catarrhine monkey), toque macaque (catarrhine monkey), Hamadryas baboon (catarrhine monkey), tufted capuchin (platyrrhine monkey) (Goto et al. 2010), aye-aye, and mangoose lemur (strepsirrhine primate) (Taufik et al. 2012). It is noteworthy that 3-FL but not 2′-FL has been detected in milk of three catarrhine monkeys including rhesus monkey, toque macaque, and Hamadryas baboon (Goto et al. 2010). Isoglobotriose has been found in the milk of only two strepsirrhine primates, viz Coquerel's sifaka and mongoose lemur, but not in greater galago and aye-aye (Taufik et al. 2012). Oligosaccharides containing Lewis a have been found in the milk of humans (Urashima et al. 2011, 2012) and aye-aye (Taufik et al. 2012).

Oligosaccharides containing Lewis x have also been found in milk of the platypus (monotreme) (Jenkins et al. 1984; Amano et al. 1985), echidna (monotreme) (Jenkins et al. 1984), koala (marsupial) (Urashima et al. 2013), Asian elephant (Uemura et al. 2006), and African elephant (Osthoff et al. 2008) (Proboscidea). Oligosaccharides containing the α-Gal epitope, including isoglobotriose, have also been detected in milk/colostrum of the cow (Artiodactyla) (Urashima et al. 1991a; Mariño et al. 2011), goat (Artiodactyla) (Urashima et al. 1994), sheep (Artiodactyla) (Urashima et al. 1989a), giant anteater (Pilosa) (Urashima et al. 2008), island flying fox (Chiroptera) (Senda et al. 2011), Asian and African elephants (Proboscidea) (Uemura et al. 2006; Osthoff et al. 2008). A-tetrasaccharide has been also found in milk of the minke whale (Cetacea) (Urashima et al. 2002).

1.7 Molecular Species of Sialic Acid in Milk Oligosaccharides

The acidic milk oligosaccharides contain either sialic acid or, in rare cases, sulfate at their nonreducing ends. The most predominant sialic acid is *N*-acetylneuraminic acid (Neu5Ac) in milk oligosaccharides and other glycoconjugates, followed by *N*-glycolylneuraminic acid (Neu5Gc). Sialyl milk oligosaccharides of most species contain only Neu5Ac or more Neu5Ac than Neu5Gc, but in a few mammalian species Neu5Gc predominates over Neu5Ac, or only Neu5Gc is present.

In humans, glycoconjugates contain only Neu5Ac with no more than traces of Neu5Gc; to date no Neu5Gc has been detected in human milk oligosaccharides. This is due to the absence in humans of CMP-Neu5Ac hydroxylase, which converts CMP-Neu5Ac to CMP-Neu5Gc (Brinkman-van den Linden et al. 2000); traces of Neu5Gc are thought to derive from the diet. In cows (Tao et al. 2008; Mariño et al. 2011), goats (Urashima et al. 1997a), sheep (Nakamura et al. 1998), and pig (Tao et al. 2010), sialyl milk oligosaccharides contain both Neu5Ac and Neu5Gc. In cows, Neu5Ac oligosaccharides predominate over those containing Neu5Gc (Tao et al. 2008; Mariño et al. 2011), whereas in sheep the reverse is the case (Nakamura et al. 1998).

Studies on the sialic acid species in milk oligosaccharides of species of Carnivora have provided interesting results. Only Neu5Ac was found in milk oligosaccharides of the striped skunk (Taufik et al. 2013), mink (Urashima et al. 2005), Japanese black bear (Urashima et al. 2004b), and seals (Urashima et al. 2003a, 2004a; Kinoshita et al. 2009), all of which are classified as Canoidea, while only Neu5Gc has been detected in milk of the African lion and clouded leopard (Senda et al. 2010), which are members of Felidae in Feloidea. However, only Neu5Ac has been found in milk oligosaccharide of the spotted hyena (Uemura et al. 2009), a hyenid species in the Feloidea.

With respect to primate sialyl oligosaccharides, both Neu5Gc and Neu5Ac have been found in the milks of great apes including chimpanzee, bonobo, gorilla, and orangutan, while only Neu5Ac was detected in the milk of the siamang, a lesser ape (Urashima et al. 2009). In both catarrhine monkeys—including rhesus monkey, toque macaque, and Hamadryas baboon—and platyrrhine monkeys—including tufted capuchin, mantled howler, and Bolivian squirrel monkey—only Neu5Ac has been found in milk oligosaccharides (Goto et al. 2010). Among strepsirrhine primates, Neu5Gc as well as Neu5Ac have been found in greater galago milk oligosaccharides, while only Neu5Ac has been detected in milk oligosaccharides of the aye-aye, Coquerel's sifaka, and mongoose lemur (Taufik et al. 2012).

Regarding milk oligosaccharides of other eutherians, only Neu5Ac has been found in those of camels (Fukuda et al. 2010; Alhaj et al. 2013), horse (Nakamura et al. 2001), minke whale (Urashima et al. 2002), Bryde's whale (Urashima et al. 2007), sei whale (Urashima et al. 2007), and bottlenose dolphin (Uemura et al. 2005), while Neu5Gc as well as Neu5Ac have been detected in the giant anteater (Urashima et al. 2008), and only Neu5Gc has been found in the island flying fox (Senda et al. 2011).

In marsupial milk oligosaccharides, only Neu5Ac has been found in the red kangaroo (Anraku et al. 2012). As noted above, echidna milk oligosaccharide uniquely contains 4-*O*-acetyl-Neu5Ac (Messer 1974; Kamerling et al. 1982).

1.8 Type I and Type II Chains in Milk Oligosaccharides

Milk oligosaccharides with core structural units other than lactose (see, for example, the core structures of human milk oligosaccharides in Fig. 1.6) contain either Gal(β1-4)GlcNAc (*N*-acetyllactosamine, LacNAc) or Gal(β1-3)GlcNAc (lacto-*N*-biose I, LNB) or both structures. Those oligosaccharides containing LNB are termed type I, while those containing LacNAc are type II. Other examples of oligosaccharides of types I and II are shown in Fig. 1.15.

The presence/absence of type I or type II, as well as their ratios in milk oligosaccharides, differ among mammalian species. For example in human milks, the concentrations of representative oligosaccharides were determined during the course of lactation; it was shown that the predominant saccharides were 2'-FL,

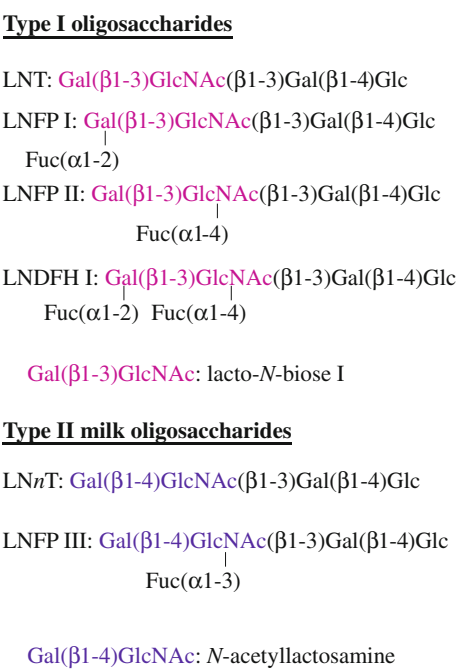


Fig. 1.15 Structures of type I and type II human milk oligosaccharides

LNFP-I, Fuc(α1-2)Gal(β1-3)[Fuc(α1-4)]GlcNAc(β1-3)Gal(β1-4)Glc (lacto-*N*-difucohexaose I, LNDFH-I), and Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)Glc (lacto-*N*-tetraose, LNT) (Chaturvedi et al. 2001; Thurl et al. 2010; Asakuma et al. 2008). It should be noted that except for 2'-FL, which is neither type I nor type II, all are type I oligosaccharides. In human milk, type I oligosaccharides predominate over type II including LN*n*T and Gal(β1-4)[Fuc(α1-3)]GlcNAc(β1-3)Gal(β1-4)Glc (lacto-*N*-fucopentaose III, LNFP III).

Among those mammalian species whose milk oligosaccharides have been characterized to date, the type II but not type I have been found in cows (Mariño et al. 2011), horse (Urashima et al. 1989b, 1991b), camels (Fukuda et al. 2010; Alhaj et al. 2013), several carnivorans (Urashima et al. 1997b, 1999a, c, 2000, 2001b, 2003a, 2004a, b, 2005; Taufik et al. 2013; Kinoshita et al. 2009), Asian and African elephants (Uemura et al. 2006; Osthoff et al. 2008), minke whale (Urashima et al. 2002), island flying fox (Senda et al. 2011), giant anteater (Urashima et al. 2008), tammar wallaby (Bradbury et al. 1983), red kangaroo (Anraku et al. 2012), koala (Urashima et al. 2013), and others. Although both types have been found in platypus milk, type II predominate over the type I (Amano et al. 1985).

The present authors studied milk oligosaccharides of several primates including apes and monkeys, and found that although both types I and II occur in the milk/colostrum of chimpanzee, bonobo, and orangutan, type II predominates over type I in these apes (Urashima et al. 2009). Only type II and not type I saccharides have been found in milk/colostrum of gorilla, siamang (Urashima et al. 2009), rhesus

monkey, toque macaque, Hamadryas baboon, tufted capuchin (Goto et al. 2010), greater galago, Coquerel's sifaka, and mangoose lemur (Taufik et al. 2012). Although both types have been detected in milk of the aye-aye, type II predominates over the type I (Taufik et al. 2012). The ratio of type I to type II oligosaccharides in milk presumably depends on the relative activities of β3 and β4 galactosyltransferases, both of which transfer Gal to GlcNAc, in lactating mammary glands (see Fig. 1.14).

As milk of the platypus, a monotreme, contains both type I and type II oligosaccharides, we hypothesize that the milk of the common ancestor of mammals similarly contained both types. Furthermore, as the type II predominates over type I in the milk/colostrum of all studied non-human species, it appears that human milk is exceptional with respect to the predominance of type I oligosaccharides.

What is the biological significance for human infants of type I predominance of milk oligosaccharides? The authors have hypothesized that this predominance may be a selective advantage in that type I, but not type II, stimulate the establishment of a bifidus flora in the colon of breast-fed infants (Urashima et al. 2012). As a result of the colonization of bifidobacteria in the infant colon, the colonic pH becomes acidic and the growth of pathogenic microorganisms is thereby inhibited.

1.9 The Evolutionary-Developmental Context within Which Milk Oligosaccharides Evolved

A recent review of the evolution of the developmental patterns of mammary glands revealed several important points (Oftedal and Dhouailly 2013). First, the mammary gland apparently evolved via incorporation of a developmental triad, the apo-pilo-sebaceous unit (APSU) into a more complex structure with deeper penetration into the dermis and subdermal adipose tissue. Second, the most primitive form of mammary gland appears to be the nipple-less mammary alveola or milk patch, in which the mammary lobules each empty via a separate duct into the infundibulum of an enlarged hair follicle, the so-called mammary hair. The entire mammary structure, consisting of 100–200 independent lobules, originates from one ectodermal placode. Third, this structure appears to have evolved in such a way as to allow dispersion of mammary secretion across a broader surface area, and thus may have initially functioned as a means to provide moisture and antimicrobials to the surface of incubated eggs. Fourth, based on molecular genetic estimates of the time of evolutionary appearance of diverse milk proteins—many of which appear to have derived from antimicrobial secretory products of a generalized skin gland—the application of protolacteal secretions to eggs not only long predates the appearance of mammals in the Jurassic, but may also predate the transition from early tetrapods to amniotes in the Carboniferous, more than 310 MYBP.

Thus, it now appears likely that the ancestral apocrine-like glands that evolved into the secretory tissue in mammary glands were originally involved in moistening eggs (Oftedal 2002), and may have been an important component of the tetrapod transition to egg-laying on land. Unlike hard-shelled turtle, crocodile, and bird eggs, the parchment-shelled eggs of lepidosaurs (lizards and snakes) and egg-laying monotremes are very susceptible to both moisture loss and microbial infection. Many extant amphibians produce a diverse array of antimicrobial constituents in their skin secretions, and it may be that such constituents developed importance in egg protection (Oftedal 2013). It was such constituents that evolved into most milk proteins, e.g., c-lysozyme into α-lactalbumin, a retinol-binding protein (lipocalin) into β-lactoglobulin, a Whey Acidic Protein Four-Disulphide Core (WFDC) protein into whey acidic protein, and xanthine oxidase into a key fat globule membrane protein (Oftedal 2013). Even caseins, which evolved from secretory calcium-binding phosphoproteins such as odontogenic ameloblast-associated protein and follicular dendritic cell-secreted peptide may have initially been concerned with regulation of calcium transfer to egg surfaces (Kawasaki et al. 2011), as eggs without calcified eggshells (an evolutionary novelty of the sauropsids, or ancestors of "reptiles" and birds) are typically calcium-poor (Oftedal 2002, 2013).

Seen in this light, the origin and initial evolution of oligosaccharides may have been associated with protection of egg and skin surfaces in a mammary area, not just colonization of the hatchling gastrointestinal tract. Once the nipple developed—after the divergence of the ancestors of monotremes and therians—the importance of oligosaccharides would have been altered to more of a gastrointestinal role. It is also likely that as egg yolk proteins diminished in importance with egg size reduction (Brawand et al. 2008), oligosaccharides may have taken on new nutritional roles, such as providing a source of preformed glucose and sialic acid to increasingly altricial (immature) offspring. For example, preformed glucose and sialic acid may have been required as substrates for developing tissues, including the brain (Smith 2006; Wang 2009; Eisert et al. 2013a, b).

Eutherians managed to evolve a series of complex placental types that could take over many of the earlier functions of mammary secretions in nourishing developmentally immature young. This likely changed the selective pressures acting on milk composition, removing constraints imposed by the functional needs of very altricial neonates. Thus, for the first time we find high-lactose, low-oligosaccharide milks, especially in eutherian taxa with precocial (developed) young at birth, such as artiodactyls and perissodactyls (see above). Fur seals and sealions (family Otariidae) also produce precocial young, but have lost the ability to secrete functional α-lactalbumin (Sharp et al. 2008) and thus to secrete either lactose or oligosaccharide in milk (Messer and Urashima 2002). This is consistent with the hypothesis that milk saccharides may not be as important for slow-growing species with precocial young (Eisert et al. 2013a). If so, some eutherians may have found alternative methods to protect the gastrointestinal tract of neonates from microbial pathogens, because the dense breeding colonies of otariids are certainly ripe locations for microbial proliferation.

It is intriguing, but not well understood, why milk oligosaccharides should have secondarily evolved such an important role in human infant health and development. Human milk oligosaccharides are not only much more diverse, but are apparently at higher concentrations than in any other primate, including great apes; they also include, apparently for the first time in eutherian evolution, a preponderance of type I structures (Urashima et al. 2009, 2012). Whether this aspect of human milk oligosaccharides has any correlation to the secondary reduction in developmental maturity of humans at birth, compared to great apes and some other primates, remains to be determined.

References

Alhaj OA, Taufik E, Handa Y, Fukuda K, Saito T, Urashima T (2013) Chemical characterization of oligosaccharides in commercially pasteurized dromedary camel (*Camelus dromedaries*) milk. Int Dairy J 28:70–75. doi:10.1016/j.idairyj.2012.08.008

Amano J, Messer M, Kobata A (1985) Structures of the oligosaccharides isolated from milk of the platypus. Glycoconj J 2:121–135

Amano J, Osanai M, Orita T, Sugawara D, Osumi K (2009) Structural determination by negative-ion MALDI-QIT-TOFMS^n after pyrene derivatization of variously fucosylated oligosaccharides with branched decaose cores from human milk. Glycobiology 19:601–614. doi:10.1093/glycob/cwp026

Anraku T, Fukuda K, Saito T, Messer M, Urashima T (2012) Chemical characterization of acidic oligosaccharides in milk of the red kangaroo (*Macropus rufus*). Glycoconj J 29:147–156. doi:10.1007/s10719-012-9372-7

Asakuma S, Urashima T, Akahori M, Ohbayashi H, Nakamura T, Kimura K, Watanabe Y, Arai I, Sanai Y (2008) Variation of major neutral oligosaccharides levels in human colostrum. Eur J Clin Nutr 62:488–494. doi:10.1038/sj.ejcn.1602738

Asakuma S, Hatakeyama E, Urashima T, Yoshida E, Katayama T, Yamamoto K, Kumagai H, Ashida H, Hirose J, Kitaoka M (2011) Physiology of consumption of human milk oligosaccharides by infant gut-associated Bifidobacteria. J Biol Chem 286:34583–34592. doi:10.1074/jbc.M111.248138

Barra WC, Caputto R (1965) Isolation and identification of a lactose sulphate ester from rat mammary glands. Biochim Biophys Acta 101:367–369

Bode L, Rudloff S, Kunz C, Strobel S, Klein N (2004) Human milk oligosaccharides reduce platelet-neutrophil complex formation leading to a decrease in neutrophil β2 integrin expression. J Leukoc Biol 76:820–826

Bode L (2006) Recent advances on structure, metabolism and function of human milk oligosaccharides. J Nutr 136:2127–2130

Bradbury JH, Collins JG, Jenkins GA, Trifonoff E, Messer M (1983) ^{13}C-NMR study of the structures of two branched oligosaccharides from marsupial milk. Carbohydr Res 122:327–331

Brawand D, Wahli W, Kaessmann H (2008) Loss of egg yolk genes in mammals and the origin of lactation and placentation. PLoS Biol 6:e63

Brinkman-van den Linden ECM, Sjoberg ER, Raj Juneja L, Crocker PR, Varki N, Varki AJ (2000) Loss of *N*-glycolylneuraminic acid in human evolution, Implication for sialic acid recognition by siglecs. J Biol Chem 275:8633–8640

Bubb WA, Urashima T, Kohso K, Nakamura T, Arai I, Saito T (1999) Occurence of an unusual lactose sulfate in dog milk. Carbohydr Res 318:123–128

Chaturvedi P, Warren CD, Altaye M, Morro AI, Ruiz-Palacios G, Pickerling IK, Newburg DS (2001) Fucosylated human milk oligosaccharides vary between individuals and over the course of lactation. Glycobiology 11:365–372

Choi HU, Carubelli R (1968) Neuraminin-lactose, neuraminin-lactose sulfate, and lactose sulfate from rat mammary glands. Isolation, purification, and permethylation studies. Biochemistry 7:4423–4430

Collins JG, Bradbury JH, Trifonoff E, Messer M (1981) Structures of four new oligosaccharides from marsupial milk, determined mainly by ^{13}C-NMR spectroscopy. Carbohydr Res 92:136–140

Crisp EA, Messer M, Cowan PE (1989a) Intestinal lactase (β-galactosidase) and other disaccharidase activities of suckling and adult common brushtail possums. *Trichosurus vulpecula* (Marsupialia: Phalangeridae). Reprod Fertil Dev 1:315–324

Crisp EA, Messer M, VandeBerg JL (1989b) Changes in milk carbohydrates during lactation in a didelphid marsupial *Monodelphis domestica*. Physiol Zool 62:1117–1125

Donald ASR, Feeney J (1988) Separation of human milk oligosaccharides by recycling chromatography. First isolation of lacto-*N*-neo-difucohexaose II and 3'-galactosyllactose from this source. Carbohydr Res 178:79–91

Eisert R, Oftedal OT, Barrell GK (2013a) Milk composition in the Weddell seal (*Leptonychotes weddellii*): evidence for a functional role of milk carbohydrates in pinnipeds. Physiol Biochem Zool 86:159–175

Eisert R, Potter C, Oftedal OT (2013b) Brain size in neonatal and adult Weddell seals: costs and consequences of having a large brain. Mar Mammal Sci. doi:10.1111/mms.12033

Engfer MB, Stahl B, Finke B, Sawatzki G, Daniel H (2000) Human milk oligosaccharides are resistant to enzymatic hydrolysis in the upper gastrointestinal tract. Am J Clin Nutr 71:1589–1596

Fukuda K, Yamamoto A, Ganzorig K, Khuukhenbaatar J, Senda A, Saito T, Urashima T (2010) Chemical characterization of the oligosaccharides in Bactrian camel (*Camelus bactrianus*) milk and colostrum. J Dairy Sci 93:5572–5587. doi:10.3168/jds.2010-3152

Gnoth MJ, Kunz C, Kinne-Saffran E, Rudloff S (2000) Human milk oligosaccharides are minimally digested in vitro. J Nutr 130:3014–3020

Goto K, Fukuda K, Senda A, Saito T, Kimura K, Glander KE, Hinde K, Dittus W, Milligan LA, Power ML, Oftedal OT, Urashima T (2010) Chemical characterization of oligosaccharides in the milk of six species of new and old world monkeys. Glycoconj J 27:703–715. doi:10.1007/s10719-010-9315-0

Grönberg G, Lipniunas P, Lundgren T, Lindh F, Nilsson B (1992) Structural analysis of five new monosialylated oligosaccharides from human milk. Arch Biochem Biophys 296:597–610

Jantscher-Krenn E, Zherebtsov M, Nissan C, Goth K, Guner YS, Naidu N, Choudbury B, Grishin AV, Ford HR, Bode L (2012) The human milk oligosaccharide disialyllacto-*N*-tetraose prevents necrotizing enterocolitis in neonatal rats. Gut 61:1417–1425. doi:10.1136/gutjnl-2011-301404

Jenkins GA, Bradbury JH, Messer M, Trifonoff E (1984) Determination of the structures of fucosyl-lactose and difucosyl-lactose from the milk of monotremes, using ^{13}C-n.m.r. spectroscopy. Carbohydr Res 126:157–161

Jenness R, Regehr EA, Sloan RE (1964) Comparative studies of milks. II. Dialyzable carbohydrates. Comp Biochem Physiol 13:339–352

Kamerling JP, Dorland L, van Halbeek H, Vliegenthart JFG, Messer M, Schauer R (1982) Structural studies of 4-*O*-acetyl-α-*N*-acetylneuraminyl-(2,3)-lactose, the main oligosaccharide in echidna milk. Carbohydr Res 100:331–340

Kawasaki K, Lafont A, Sire J (2011) The evolution of milk casein genes from tooth genes before the origin of mammals. Mol Biol Evol 28:2053–2061

Kinoshita M, Ohta H, Higaki K, Kojima Y, Urashima T, Nakajima K, Suzuki M, Kovacs KM, Lydersen C, Hayakawa T, Kakehi K (2009) Structural characterization of multi-branched oligosaccharides from seal milk by combination of off-line HPLC-MALDI-TOF MS and sequential exoglycosidase digestion. Anal Biochem 388:242–253

Klein A, Schwertman A, Peters M, Kunz C, Strobel S (2000) Immunomodulatory effects of breast milk oligosaccharides. In: Koletzki B (ed) Short and long term effects of breast feeding on child health. Kluwer Academic/Plenum Publishers, New York, pp 251–259

Kobata A (2010) Structures and application of oligosaccharides in human milk. Proc Jpn Acad Ser B Phys Biol Sci 86:1–17

Luo Z-X, Yuan C-X, Meng Q-J, Ji Q (2011) A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476:442–445

Mariño K, Lane JA, Abrahams JL, Struwe WB, Harvey DJ, Marotta M, Hickey RM, Rudd PM (2011) Method for milk oligosaccharide profiling by 2-aminobenzamide labeling and hydrophilic interaction chromatography. Glycobiology 21:1317–1330. doi:10.1093/glycob/cwr067

McKenzie HA, White FH Jr (1991) Lysozyme and α-lactalbumin, structure, function, and interrelationships. Adv Protein Chem 41:174–315

Messer M, Kerry KR (1973) Milk carbohydrates of the echidna and the platypus. Science 180:201–203

Messer M (1974) Identification of *N*-acetyl-4-*O*-acetylneuraminyl-lactose in echidna milk. Biochem J 139:415–420

Messer M, Green B (1979) Milk carbohydrates of marsupials II. Quantitative and qualitative changes in milk carbohydrates during lactation in the tammar wallaby (*Macropus eugenii*). Aust J Biol Sci 32:519–531

Messer M, Trifonoff E, Stern W, Collins JG, Bradbury JH (1980) Structure of a marsupial milk trisaccharide. Carbohydr Res 83:327–334

Messer M, Trifonoff E, Collins JG, Bradbury JH (1982) Structure of a branched tetrasaccharide from marsupial milk. Carbohydr Res 102:316–320

Messer M, Gadiel PA, Ralston GB, Griffiths M (1983) Carbohydrates of the milk of the platypus. Aust J Biol Sci 36:129–137

Messer M, Crisp EA, Czolij R (1989) Lactose digestion in suckling macropodids. In: Grigg G, Jarman P, Hume I (eds) Kangaroos, Wallabies and Rat Kangaroos. Surry Beatty & Sons Pty Ltd, Australia, pp 217–221

Messer M, Nicholas KR (1991) Biosynthesis of marsupial milk oligosaccharides: characterization and developmental changes of two galactosyltransferases in lactating mammary glands of the tammar wallaby, *Macropus eugenii*. Biochim Biophys Acta 1077:79–85

Messer M, Griffiths M, Rismiller PD, Shaw BC (1997) Lactose synthesis in a monotreme, the echidna (*Tachyglossus aculeatus*): isolation and amino acid sequence of echidna α-lactalbumin. Comp Biochem Physiol B118:403–410

Messer M, Weiss AS, Shaw DC, Westerman MJ (1998) Evolution of the monotremes: phylogenetic relationship to marsupials and eutherians, and divergence dates based on α-lactalbumin amino acid sequences. J Mammalogy 5:95–105

Messer M, Urashima T (2002) Evolution of milk oligosaccharides and lactose. Trends Glycosci Glycotech 14:153–176

Nakamura T, Urashima T, Nakagawa M, Saito T (1998) Sialyllactose occurs as free lactones in ovine colostrum. Biochim Biophys Acta 1381:286–292

Nakamura T, Amikawa S, Harada T, Saito T, Arai I, Urashima T (2001) Occurrence of an unusual phosphorylated *N*-acetyllactosamine in horse colostrum. Biochim Biophys Acta 1525:13–18

Nakamura T, Urashima T, Mizukami T, Fukushima M, Arai I, Senshu T, Imazu K, Nakao T, Saito T, Ye Z, Zuo H, Wu K (2003) Composition and oligosaccharides of a milk sample of the giant panda, *Ailuropoda melanoleuca*. Comp Biochem Physiol B135:439–448. doi:10.1016/S1096-4959(03)00093-9

Newburg DS, Naubauer SH (1995) Carbohydrate in milks: analysis, quantities and significance. In: Jensen RG (ed) Handbook of milk compositions. Academic Press, San Diego, pp 273–349

Ninonuevo MR, Park Y, Yin H, Zhang J, Ward RE, Clowers BH, German JB, Freeman SL, Killeen K, Grimm R, Lebrilla CB (2006) Strategy for annotating the human milk glycome. J Agric Food Chem 54:7471–7480

Obermeier S, Rudloff S, Pohlentz S, Lentze MJ, Kunz C (1999) Secretion of ^{13}C-labelled oligosaccharides into human milk and infant's urine after an oral [^{13}C] galactose load. Isotopes Environ Health Stud 35:119–125

Oftedal OT (2002) The origin of lactation as a water source for parchment-shelled eggs. J Mammary Gland Biol 7:253–266

Oftedal OT (2013) Origin and evolution of the major constituents of milk. In: McSweeney PHL, Fox PF (eds) Advanced dairy chemistry, vol 1A, 4th edn. Springer Science + Buisiness Media, New York, pp 1–42

Oftedal OT, Dhouailly D (2013) Evo-Devo of the mammary gland. J Mammary Gland Biol Neoplasia 18:105–120

Osthoff G, Dickens L, Urashima T, Bonnet SL, Uemura Y, van der Westhuizen JH (2008) Structural characterization of oligosaccharides in the milk of an African elephant (*Loxodonta africana africana*). Comp Biochem Physiol B150:74–84. doi:10.1016/j/cbpb.2008.01.010

Rudloff S, Pohlentz G, Diekman L, Egge H, Kunz C (1996) Urinary excretion of lactose and oligosaccharides in preterm infants fed human milk or infant formula. Acta Paediatr 85:598–603

Ruiz-Palacios GM, Cervantes LE, Romas P, Chavez-Munguia B, Newburg DS (2003) *Campylobacter jejuni* binds intestinal H(O) antigen (Fucα1,2Galβ1,4GlcNAc), and fucosyloligosaccharides of human milk inhibit its binding and infection. J Biol Chem 278:14112–14120

Saito T, Itoh T, Adachi S (1984) Presence of two neutral disaccharides containing *N*-acetylhexosamine in bovine colostrum as free forms. Biochim Biophys Acta 801:147–150

Saito T, Itoh T, Adachi S (1987) Chemical structures of three neutral trisaccharides isolated in free forms from bovine colostrum. Carbohydr Res 165:43–51

Senda A, Hatakeyama E, Kobayashi R, Fukuda K, Uemura Y, Saito T, Packer C, Oftedal OT, Urashima T (2010) Chemical characterization of milk oligosaccharides of an African lion (*Panthera leo*) and a clouded leopard (*Neofelis nebulosa*). Anim Sci J 81:687–693. doi:10.1111/j.1740-0929.2010.00787.x

Senda A, Kobayashi R, Fukuda K, Saito T, Hood WR, Kunz TH, Oftedal OT, Urashima T (2011) Chemical characterization of milk oligosaccharides of the island flying fox (*Pteropus hypomelanus*) (Chiroptera: Pteropodidae). Anim Sci J 82:782–786. doi:10.1111/j.1740-0929.2011.00906.x

Shaper NL, Charron M, Lo NW, Shaper JH (1998) β-1,4-galactosyltransferase and lactose biosynthesis: recruitment of a housekeeping gene from the nonmammalian vertebrate gene pool for a mammary gland specific function. J Mammary Gland Biol 3:315–324

Sharp JA, Lefèvre C, Nicholas KR (2008) Lack of functional alpha-lactalbumin prevents involution in Cape fur seals and identifies the protein as an apoptotic milk factor in mammary gland involution. BMC Biol 6:48

Shaw DC, Messer M, Scrivener AM, Nicholas KR, Griffiths M (1993) Isolation, partial characterization, and amino acid sequence of α-lactalbumin from platypus (*Ornithorhynchus anatinus*) milk. Biochim Biophys Acta 1161:177–186

Smith KK (2006) Craniofacial development in marsupial mammals: developmental origins of evolutionary change. Dev Dyn 235:1181–1193

Stewart IM, Messer M, Walcott PJ, Gadiel PA, Griffiths M (1983) Intestinal glycosidase activities in one adult and two suckling echidnas: absence of a neutral lactase (β-D-galactosidase). Aust J Biol Sci 36:139–146

Sturman JA, Lin YY, Higuchi T, Fellman JH (1985) *N*-Acetylneuramin lactose sulfate: a newly identified nutrient in milk. Pediatr Res 19:216–219

Tao N, DePeters EJ, Freeman S, German JB, Grimm R, Lebrilla CB (2008) Bovine milk glycome. J Dairy Sci 91:3768–3778

Tao N, Ochonicky KL, German JB, Donovan SM, Lebrilla CB (2010) Structural determination and daily variations of porcine milk oligosaccharides. J Agric Food Chem 58:4653–4659

Tao N, Wu S, Kim J, An HJ, Hinde K, Power ML, Gangneux P, German JB, Lebrilla CB (2011) Evolutionary glycomics: characterization of milk oligosaccharides in primates. J Proteome Res 10:1548–1557

Taufik E, Fukuda K, Senda A, Saito T, Williams C, Tilden C, Eisert R, Oftedal OT, Urashima T (2012) Structural characterization of neutral and acidic oligosaccharides in the milks of strepsirrhine primates: greater galago, aye-aye, Coquerel's sifaka and mongoose lemur. Glycoconj J 29:119–134. doi:10.1007/s10719-012-9370-9

Taufik E, Sekii N, Senda A, Fukuda K, Saito T, Eisert R, Oftedal OT, Urashima T (2013) Neutral and acidic milk oligosaccharides of the striped skunk (Mephitidae: *Mephitis mephitis*). Anim Sci J 84:569–578. doi:10.1111/asj.12040

Taufik E, Ganzorig K, Nansalmaa M, Fukuda R, Fukuda K, Saito T, Urashima T (2014) Chemical characterization of saccharides in the milk of a reindeer (*Rangifer tarandus tarandus*). Int Dairy J 34:104–108. doi:10.1016/j.idairyj.2013.07.012

Thurl S, Munzert M, Henker J, Boehm G, Muller-Werner B, Jelinek J, Stahl B (2010) Variation of human milk oligosaccharides in relation to milk groups and lactation periods. Br J Nutr 104:1261–1271

Uemura Y, Asakuma S, Nakamura T, Arai I, Taki M, Urashima T (2005) Occurrence of a unique sialyl tetrasaccharide in colostrum of a bottlenose dolphin (*Tursiops truncatus*). Biochim Biophys Acta 1725:290–297. doi:10.1016/j.bbagem.2005.05.011

Uemura Y, Asakuma S, Yon L, Saito T, Fukuda K, Arai I, Urashima T (2006) Structural determination of the oligosaccharides in the milk of an Asian elephant (*Elephas maximus*). Comp Biochem Physiol A145:468–478. doi:10.1016/j.cbpa.2006.08.001

Uemura Y, Takahashi S, Senda A, Fukuda K, Saito T, Oftedal OT, Urashima T (2009) Chemical characterization of milk oligosaccharides of a spotted hyena (*crocuta crocuta*). Comp Biochem Physiol A152:158–161. doi:10.1016/j.cbpa.2008.09.013

Urashima T, Saito T, Nishimura J, Ariga H (1989a) New galactosyllactose containing α-glycosidic linkage isolated from ovine (*Booroola dorset*) colostrum. Biochim Biophys Acta 992:375–378

Urashima T, Sakamoto T, Ariga H, Saito T (1989b) Structure determination of three neutral oligosaccharides obtained from horse colostrum. Carbohydr Res 194:280–287

Urashima T, Saito T, Ohmisya K, Shimazaki K (1991a) Structural determination of three neutral oligosaccharides in bovine (Holstein-Friesian) colostrum, including the novel trisaccharide; GalNAcα1-3Galβ1-4Glc. Biochim Biophys Acta 1073:225–229

Urashima T, Saito T, Kimura T (1991b) Chemical structures of three neutral oligosaccharides obtained from horse (Thoroughbred) colostrum. Comp Biochem Physiol B100:177–183

Urashima T, Messer M, Bubb WA (1992) Biosynthesis of marsupial milk oligosaccharides II: characterization of a β^6-*N*-acetylglucosaminyltransferase in lactating mammary glands of the tammar wallaby, *Macropus eugenii*. Biochim Biophys Acta 1117:223–231

Urashima T, Bubb WA, Messer M, Tsuji Y, Taneda Y (1994) Studies of the neutral trisaccharides of goat (*Capra hircus*) colostrum and of the one- and two-dimensional ^{1}H and ^{13}C NMR spectra of 6'-*N*-acetylglucosaminyllactose. Carbohydr Res 262:173–184

Urashima T, Murata S, Nakamura T (1997a) Structural determination of monosialyl trisaccharides obtained from caprine colostrum. Comp Biochem Physiol B116:431–435

Urashima T, Kusaka Y, Nakamura T, Saito T, Maeda N, Messer M (1997b) Chemical characterization of milk oligosaccharides of the brown bear, *Ursus arctos yesoensis*. Biochim Biophys Acta 1334:247–255

Urashima T, Yamamoto M, Nakamura T, Arai I, Saito T, Namiki M, Yamaoka K, Kawahara K (1999a) Chemical characterisation of the oligosaccharides in a sample of milk of a white-nosed coati, *Nasua narica* (Procyonidae: Carnivora). Comp Biochem Physiol A123:187–193

Urashima T, Kawai Y, Nakamura T, Arai I, Saito T, Namiki M, Yamaoka K, Kawahara K, Messer M (1999b) Chemical characterization of six oligosaccharides in a sample of colostrum of the brown capuchin, *Cebus apella* (Cebidae: Primate). Comp Biochem Physiol C124:295–300

Urashima T, Sumiyoshi W, Nakamura T, Arai I, Saito T, Komatsu T, Tsubota T (1999c) Chemical characterization of milk oligosaccharides of the Japanese black bear, *Ursus thibetanus japonicus*. Biochim Biophys Acta 1472:290–306

Urashima T, Yamashita T, Nakamura T, Arai I, Saito T, Derocher AE, Wiig O (2000) Chemical characterization of milk oligosaccharides of the polar bear, *Ursus maritimus*. Biochim Biophys Acta 1475:395–408

Urashima T, Saito T, Nakamura T, Messer M (2001a) Oligosaccharides of milk and colostrum in non-human mammals. Glycoconj J 18:357–371

Urashima T, Arita M, Yoshida M, Nakamura T, Arai I, Saito T, Arnould JPY, Kovacs KM, Lydersen C (2001b) Chemical characterization of the oligosaccharides in hooded seal. (*Cystophora cristata*) and Australian fur seal (*Arctocephalus pusillus doriferus*) milk. Comp Biochem Physiol B128:307–323

Urashima T, Sato H, Munakata J, Nakamura T, Arai I, Saito T, Tetsuka M, Fukui Y, Ishikawa H, Lydersen C, Kovacs KM (2002) Chemical characterization of oligosaccharides in beluga (*Delphinapterus leucas*) and minke whale (*Balaenoptera acutorostrata*) milk. Comp Biochem Physiol B132:611–624

Urashima T, Nakamura T, Yamaguchi K, Munakata J, Arai I, Saito T, Lydersen C, Kovacs KM (2003a) Chemical characterization of the oligosaccharides in milk of high Arctic harbour seal (*Phoca vitulina vitulina*). Comp Biochem Physiol A135:549–563. doi:10.1016/S1095-6433(03)00130-2

Urashima T, Nagata H, Nakamura T, Arai I, Saito T, Imazu K, Hayashi T, Derocher AE, Wiig O (2003b) Differences in oligosaccharide pattern of a sample of polar bear colostrum and mid-lactation milk. Comp Biochem Physiol B136:887–896. doi:10.1016/jcbpc.2003.09.001

Urashima T, Nakamura T, Nakagawa D, Noda M, Arai I, Saito T, Lydersen C, Kovacs KM (2004a) Characterization of oligosaccharides in a milk of bearded seal (*Erignathus barbatus*). Comp Biochem Physiol B138:1–18. doi:10.1016/j.cbpc.2003.12.009

Urashima T, Nakamura T, Teramoto K, Arai I, Saito T, Komatsu T, Tsubota T (2004b) Chemical characterization of sialyl oligosaccharides in milk of the Japanese black bear, *Ursus thibetanus japonicus*. Comp Biochem Physiol B139:587–595. doi:10.1016/j.cbpc.2004.07.012

Urashima T, Nakamura T, Ikeda A, Asakuma S, Arai I, Saito T, Oftedal OT (2005) Characterization of oligosaccharides in milk of a mink, *Mustela vison*. Comp Biochem Physiol A142:461–471. doi:10.1016/jcbpa.2005.09.015

Urashima T, Kobayashi M, Asakuma S, Uremura Y, Arai I, Fukuda K, Saito T, Mogoe T, Ishikawa H, Fukui Y (2007) Chemical characterization of the oligosaccharides in Bryde's whale (*Balaenoptera edeni*) and sei whale (*Balaenoptera borealis lesson*) milk. Comp Biochem Physiol B146:153–159

Urashima T, Komoda M, Asakuma S, Uemura Y, Fukuda K, Saito T, Oftedal OT (2008) Structural determination of the oligosaccharides in the milk of a giant anteater (*Myrmecophaga tridatyla*). Anim Sci J 79:699–709. doi:10.1111/j.1740-0929.2008.00583.x

Urashima T, Odaka G, Asakuma S, Uemura Y, Goto K, Senda A, Saito T, Fukuda K, Messer M, Oftedal OT (2009) Chemical characterization of oligosaccharides in chimpanzee, bonobo, gorilla, orangutan and siamang milk or colostrum. Glycobiology 19:499–508. doi:10.1093/glycob/cwp006

Urashima T, Kitaoka M, Terabayashi T, Fukuda K, Ohnishi M, Kobata A (2011) Milk oligosaccharides. In: Gordon NS (ed) Oligosaccharides: sources properties and applications. Nova Science, New York, pp 1–58

Urashima T, Asakuma S, Leo F, Fukuda K, Messer M, Oftedal OT (2012) The predominance of type I oligosaccharides is a feature specific to human breast milk. Adv Nutr 3:473S–482S

Urashima T, Taufik E, Fukuda R, Nakamura T, Fukuda K, Saito T, Messer M (2013) Chemical characterization of milk oligosaccharides of the koala *(Phascolarctos cinereus*). Glycoconj J 30:801–811. doi:10.1007/s10719-013-9484-8

Walcott PJ, Messer M (1980) Intestinal lactase (β-galactosidase) and other glycosidase activities in suckling and adult tammar wallabies (*Macropus eugenii*). Aust J Biol Chem 33:521–530

Wang B (2009) Sialic acid is an essential nutrient for brain development and cognition. Annu Rev Nutr 29:177–222

Wilson DE, Reeder DM (eds) (2005) Mammal species of the world: a taxonomic and geographic reference, 3rd edn. Johns Hopkins University Press, Baltimore, p 2142

Chapter 2
Genomics-Based Insights into the Evolution of Secondary Metabolite Biosynthesis in Actinomycete Bacteria

Sergey B. Zotchev

Abstract Actinomycete bacteria are known for their ability to produce chemically diverse secondary metabolites with various biological activities, some of which are being used in human therapy as anti-microbial and anti-cancer agents. Recent genome sequencing and analyses revealed that these bacteria have a much larger potential to biosynthesize secondary metabolites that was previously assumed from a conventional bioactivity screening. Indeed, each actinomycete genome was shown to carry 20–30 gene clusters for biosynthesis of secondary metabolites, most of which are not expressed in the laboratory conditions. Detailed analysis of such gene clusters along with comparative genomics studies identified some interesting features reflecting evolution of the clusters upon transfer to a new host. Moreover, insights into the process of forming new, hybrid gene clusters via recombination events at the ends of the linear chromosomes have been gained. This chapter presents and discusses recent advances in genomics of actinomycetes and its impact on our understanding of secondary metabolism evolution in these bacteria.

2.1 Introduction

Bacteria are the most versatile living organisms on Earth, capable of occupying very diverse environmental niches, such as soil, plants and animals, glaciers, hot springs, marine, desert habitats, etc. The ability of bacteria to adapt to a particular environment is unprecedented, owing mainly to their fast mutation rate (e.g., due to SOS-response (Galhardo et al. 2007), flexible regulatory networks linked to the environmental sensors, and efficient acquisition of new genes via horizontal gene

S. B. Zotchev (✉)
Department of Biotechnology, Norwegian University of Science and Technology, 7491 Trondheim, Norway
e-mail: sergey.zotchev@ntnu.no

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_2,

transfer (Syvanen 2012)). It is well documented that exposure of a particular bacterial species to a new environment drastically changes the gene expression pattern, leading to induction of genes that may be necessary for survival and proliferation under new conditions (Dufour and Donohue 2012). However, a mere survival is not enough for proliferation of the species, and acquisition of new "beneficial" genes from bacteria that dwell and thrive in this environment becomes the next step in the adaptation process. It has also been reported that some genes, no longer required in the new environment, are lost during the adaptation, the process known as "genetic drift" responsible for drastic reduction of genome size in symbiotic bacteria (Lawrence and Hendrickson 2005).

A wide variety of bacteria and fungi are capable of synthesizing secondary metabolites, compounds that are not required for growth, but may give their producers a certain advantage. The latter can be manifested, for example, by antibiotic activity of secondary metabolites against potential competitors for nutritional sources, which occupy the same ecological niche (Challis and Hopwood 2003). Certain secondary metabolites were shown to function as signal molecules, providing means for communication between microbial cells (Yim et al. 2007). Yet other secondary metabolites serve as molecular scavengers binding metal ions (Miethke 2013). A role for secondary metabolites as signalling molecules has also been proposed after discovering their effect on gene expression of other bacteria exposed to subinhibitory concentrations of the compounds (Subrt et al. 2011). Whatever the true biological role of secondary metabolites in nature, they have attracted our attention mostly because of their biological activities. The first microbial antibiotic penicillin described by A. Flemming in 1929, was produced by a fungus *Penicillium* (Abraham 1980). This important milestone in medical science prompted intensive search for new antibiotics produced by soil-dwelling bacteria and fungi. In 1940–1970, this search had overwhelming success, leading to the discovery of many important anti-microbial and anti-cancer drugs. For example, anti-fungal antibiotic amphotericin B, anti-bacterials erythromycin and vancomycin, and anti-cancer drug daunorubicin were all discovered during this period (Fig. 2.1).

The success of microbial secondary metabolites as drugs is based on their unique properties evolved over millions of years to provide efficient interaction with biological targets in living organisms (e.g., inhibiting specific enzymes, binding DNA or RNA, or preventing protein-protein interactions). Unfortunately, irresponsible use of antibiotics has led in recent years to the emergence of pathogens resistant to nearly all existing drugs. Development of resistance to anti-cancer drugs by malignant tumor cells has also been observed. These alarming tendencies prompt immediate action towards discovery and development of novel drugs based on bioactive secondary metabolites. One of possible routs could be the use of modern recombinant DNA technologies, including synthetic biology, to create hybrid molecules that can overcome resistance mechanisms, or possess dual mode of action (Flatman et al. 2005).

Having an insight into the evolution of secondary metabolite biosynthesis could provide clues on how to rationally engineer hybrid secondary metabolites. Recent

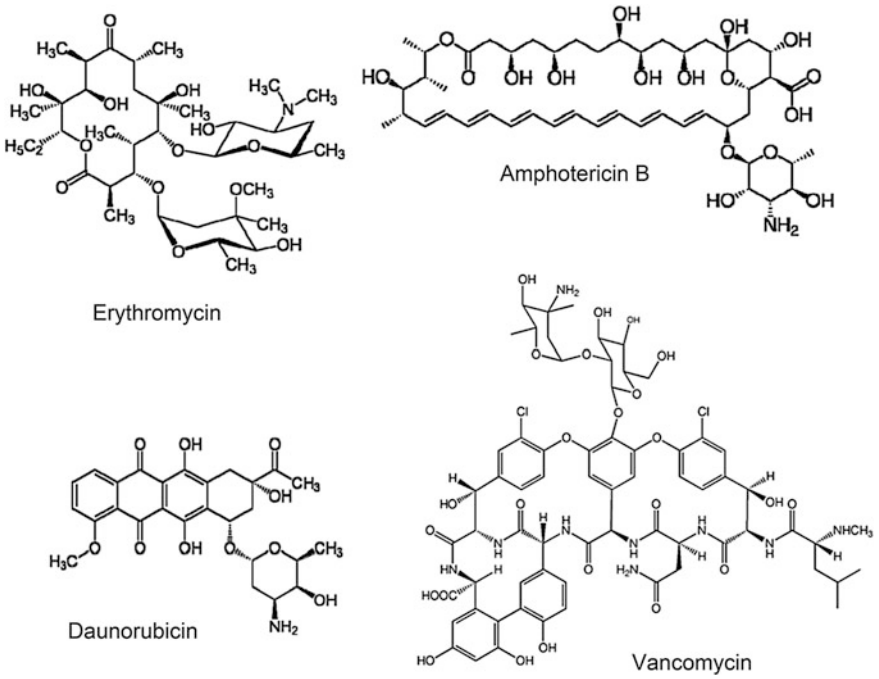


Fig. 2.1 Antibiotics and anti-cancer drugs from actinomycete bacteria

advances in genomics of microbial producers of bioactive secondary metabolites provide unique opportunity for such an investigation, which is the focus of this chapter.

2.2 Actinomycete Bacteria

Actinomycetes are filamentous Gram-positive bacteria of the order *Actinomycetales* found in a variety of diverse environmental niches, where they are usually associated with solid substrates. On solid growth medium these bacteria have a complex life cycle (Chater 1993), starting from a spore that germinates to give rise to a colony-forming mycelium, which invades the substrate and releases hydrolytic enzymes. Aerial mycelium is formed next, protruding from the surface of the colony and forming spores at the tips via septation.

These bacteria are well known for their capacity to biosynthesize biologically active secondary metabolites, which may have several functions in nature. Whatever their role is, it must be very important for actinomycetes, since genome sequencing reveals the presence of 20–30 gene clusters (most of them "silent") dedicated to the biosynthesis of secondary metabolites in each species (Doroghazi

and Metcalf 2013). It seems plausible that secondary metabolites are also important for the adaptation to new environment, as suggested by recent genome analyses of marine actinomycete of the genus *Salinispora* (Penn and Jensen 2012).

2.3 Secondary Metabolite Biosynthesis Gene Clusters and Their Dispersal

Biosynthesis of secondary metabolites (SM), and in particular antibiotics, by the actinomycete bacteria is a complex process typically governed by 10–30 genes. These genes are organized as clusters in the genomes of actinomycetes, allowing coordinated expression of the genes involved in biosynthesis, resistance and efflux of SMs. Typical SM biosynthesis gene cluster is shown in Fig. 2.2.

Biosynthetic machinery for SMs utilizes precursors from primary metabolism (e.g., amino acids, acyl-CoAs) to first build a molecular skeleton with the help of scaffold-synthesizing enzymes. Expression of genes for such enzymes is usually regulated by a positive regulator that responds to particular environmental signals.

Scaffold modification enzymes encoded by distinct genes in the cluster add chemical groups such as sugars, hydroxy, methyl, amino groups, etc., to the scaffold, thereby affording its full biological functionality. The complete molecule is biologically active, and can potentially be harmful to the producing bacterium. To avoid self-toxicity, several resistance mechanisms are usually implemented. One typically depends on the active efflux of the metabolite by a specific transporter encoded by a gene repressed by a negative regulator. The latter is only active in the absence of the molecule to be transported out of the cell, but becomes inactivated upon binding to it. Inactivation of the repressor switches on expression of the transporter, thereby ensuring efflux of SM. The second mechanism often employs a gene encoding resistance protein, which can be enzyme modifying molecular target of SM, or modifying the SM molecule itself, thereby rendering it inactive.

Interestingly, it has been shown that certain secondary metabolite biosynthesis gene clusters can be found in diverse bacterial species dwelling in geographically distinct locations (Green et al. 2008). It seems plausible that this phenomenon is due to a horizontal gene transfer (HGT) after accidental delivery of bacterial species from one location to another (e.g., by birds or fish). However, whether the same gene cluster is functional (i.e., genes expressed and compound produced) in various hosts remains unclear. In several cases it has been observed that the same, seemingly intact gene cluster for antibiotic biosynthesis is functional in one actinomycete species, while being silent in another (Zotchev unpublished data).

The method of HGT used by actinomycete bacteria to disperse SM gene clusters seems to be conjugation involving giant linear plasmids, GLPs (Kinashi et al. 1987) frequently found in actinomycetes, especially those belonging to the order *Streptomyces*. Some of these mobile genetic elements were shown to harbor single complete SM gene clusters, while others can carry several of them (Kinashi 2011).

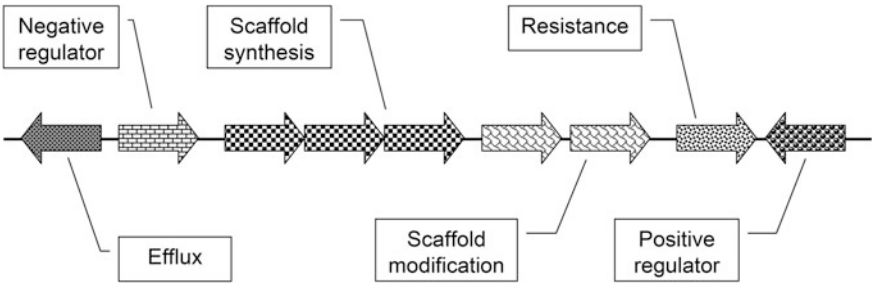


Fig. 2.2 Organization of a typical secondary metabolite biosynthesis gene cluster

The actinomycete GLPs also usually carry several IS elements and transposons, copies of which can also be found on chromosomes. It is conceivable that such elements play an important role in gene exchange between a GLP and a chromosome, thus delivering SM gene clusters to a more genetically stable location.

One of the best studied examples of a GLP carrying SM biosynthetic gene cluster is SCP1 plasmid of *Streptomyces coelicolor* A3(2). This mobile genetic element is 356 kb in size and has 75-kb terminal inverted repeats (TIRs) (Bentley et al. 2004). SCP1 is self-transmissible, and was shown to carry a complete set of genes for the biosynthesis of antibiotic methylenomycin, which can be transferred together with the plasmid to another host during conjugation (Kirby and Hopwood 1977). Recently, it has been demonstrated that SCP1 can interact with the linear chromosome of the host, forming superficial circles through interaction between terminal proteins capping ends of both linear replicons (Tsai et al. 2011). It has also been suggested that such interaction triggers recombination between GLP and chromosome to resolve post-replicational complication resulting from the formation of a superficial circle. Taking all the above into consideration, it seems plausible that GLPs play a major role in spreading the SM biosynthesis gene clusters among actinomycete bacteria, and their integration into the chromosomes.

2.4 Evolution of Secondary Metabolite Biosynthesis Pathways

In order to better understand the evolution of SM biosynthesis, one must remember that proteins participating in their biosynthetic pathways are most likely evolved from the enzymes of primary metabolism. Indeed, scaffold-building enzymes utilize precursors from primary metabolism, linking them in a particular fashion to afford a complex molecule that is distinct from a primary metabolite. Polyketide synthases type I (PKSI), enzymes governing biosynthesis of various biologically active SMs, represent a clear example of evolution of SM biosynthesis from a primary anabolic pathway. These enzymes are composed of several modules, each

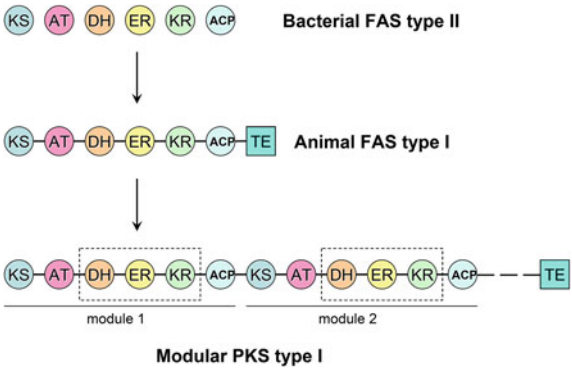


Fig. 2.3 Possible evolution of polyketide synthases type I from a fatty acid synthase. *KS* ketoacyl synthase domain, *AT* acyltransferase domain, *DH* dehydratase domain, *ER* enoyl reductase domain, *KR* ketoreductase domain; *ACP* acyl carrier protein; *TE* thioesterase domain

performing condensation of two acyl building blocks to assemble a polyketide chain. Following condensation, several reductions on the β-carbon of the polyketide chain can be catalyzed by reductive domains embedded within the module. Both the structural organization of PKS I, condensation and reduction reactions that these proteins catalyze are highly similar to an animal fatty acid synthase. The latter has most likely evolved from a bacterial type fatty acid synthase, which is not organized in modules, but is represented by discrete catalytic domains that form a multi-protein complex (Hertweck et al. 2007). A possible scheme for evolution of type I PKS is shown in Fig. 2.3. PKSI are, however, distinct from the animal type fatty acid synthase in that their modules may or may not contain all the reductive domains. Consequently, unlike fatty acids, polyketides synthesized by PKSI may have various chemical groups on their acyl chains, e.g., keto and hydroxyl groups. This chemical variability of the product is programmed into the PKSI via its structural features—number of modules, specificity for incorporated acyl units, and presence or absence of reductive domains in the modules (Cane 2010).

The diversity of SM biosynthetic pathways that must be encoded by actinomycetes genomes is astounding. So far, ca. 10,100 chemically distinct bioactive SMs from actinomycete bacteria have been reported (Mahajan and Balachandran 2012). Considering the fact that actinomycete genomes on average harbour 20–30 SM biosynthesis gene clusters, while only up to 4 of them are expressed, the number of different gene clusters (even considering redundancy) must exceed 100,000. How such diversity has been achieved during the evolution is a mystery, but some insights into the process can be gained via studying structural organization of clusters and phylogenetic analysis of their genes. Jørgensen et al. (2010) cloned a biosynthetic gene cluster for the biosynthesis of a cytotoxic macrolactam ML-449 structurally similar to the earlier reported compound BE-14106 (Jørgensen et al. 2009). Comparison of biosynthetic gene clusters for ML-449 and BE-14106 clearly indicated that the latter cluster undergone an in-frame deletion affecting PKS responsible for the biosynthesis of an aminoacyl precursor. This has resulted in reduction of length of the precursor acyl chain by two carbon atoms,

substantially increasing the cytotoxic activity of BE-14106 compared to ML-449. Moreover, detailed phylogenetic analysis of the KS domains of the PKS proteins involved in the biosynthesis of macrolactams suggested distinct evolutionary origins of the enzymes responsible for assembly of different parts of the molecules. Apparently, convergence of several biosynthetic pathways during evolution resulted in a hybrid pathway leading to production of these macrolactams.

Another example of pathway convergence leading to a hybrid SM is represented by simocyclinone D8, an inhibitor of bacterial DNA gyrase with dual mode of binding to the target (Sissi et al. 2010). The simocyclinone D8 molecule consists of several distinct chemical moieties: aminocoumarine, polyene acyl chain, deoxysugar, and angucycline. Cloning and analysis of the simocyclinone D8 gene cluster revealed the presence of four subclusters, each responsible for the biosynthesis of a separate moiety, which are joined together by specialized enzymes (Trefzer et al. 2002). Unfortunately, no phylogenetic analysis on this cluster which could reveal evolutionary origins of the sub-clusters has been reported.

2.5 Clues on the Evolution of Secondary Metabolism from Comparative Genomics

Recent advances in genome sequencing and bioinformatic analysis revolutionized our view on bacterial evolution. Comparative genomics allows to trace recent loss and acquisition of specific genes, and to distinguish between vertical and horizontal gene transfer (VGT and HGT). Actinomycete bacteria possess next largest (after myxobacteria) genomes in the bacterial domain of prokaryotes, ranging in size from ca. 5 to 11 Mb and represented by both linear (e.g., *Streptomyces*) and circular (e.g., *Salinispora*) replicons (Zhou et al. 2011; Udwary et al. 2007). Recent study by Doroghazi and Metcalf (2013) compared 102 complete genomes of actinomycetes belonging to six different genera: *Mycobacterium*, *Corynebacterium*, *Rhodococcus*, *Arthrobacter*, *Frankia*, and *Streptomyces*, focussing of SM biosynthetic gene clusters. It has been found that certain SM gene clusters are genus—specifically conserved within *Mycobacterium*, *Streptomyces,* and *Frankia*, suggesting that these clusters and corresponding SMs may play an important ecological role for these organisms.

From this point of view, it is interesting to take a closer look at the SM biosynthesis gene clusters in the genomes of actinomycetes occupying the same environmental niche. We have recently isolated several *Streptomyces* spp. from a marine sponge *Antho dichotoma* collected at the bottom of the Trondheim fjord (Norway). Draft genome sequencing of 4 phylogenetically distinct (16S RNA sequence identity <95 %) isolates and analysis of the genomes using online software antiSMASH (Blin et al. 2013) revealed gene clusters that are common to certain isolates (Zotchev, unpublished data). For example, genomes of *Streptomyces* sp. MP99-11 and *Streptomyces* sp. 115-17 share 2 clusters for the biosynthesis of

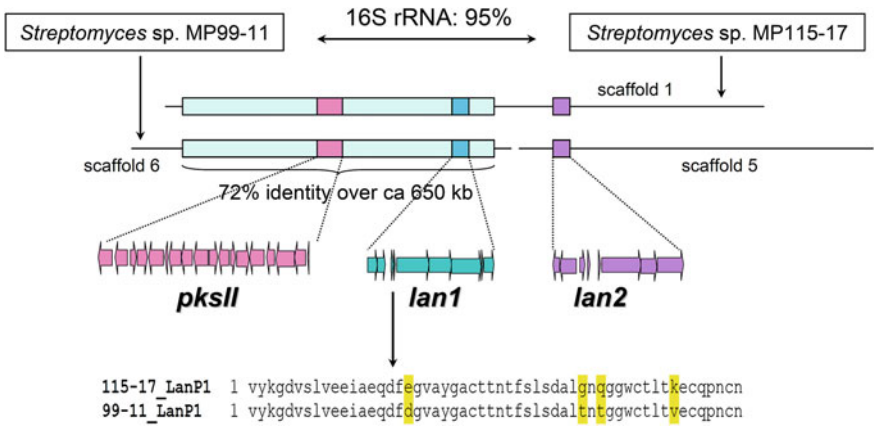


Fig. 2.4 Alignment of genomic regions of *Streptomyces* sp. MP99-11 and *Streptomyces* sp. MP115-17 harbouring gene clusters for aromatic polyketide (*pksII*) and lantibiotics (*lan1* and *lan2*). Alignment of the prepropeptides encoded by the cluster *lan1* in two genomes is shown

lantibiotics (ribosomally synthesized and post-translationally modified peptides) and an aromatic polyketide (synthesized by type II PKS). Alignment of the genomic regions harboring these cluster is presented in Fig. 2.4. Interestingly, both *pksII* and *lan1* gene clusters were co-localized within a ca 650-kb genomic region that shares overall 72 % identity on the nucleotide level. Such high degree of conservation over a large DNA fragment clearly suggests that it has been inherited by both isolates from a common ancestor, i.e., via VGT. We also observed a functional evolution of the *lan1* cluster, where the genes for prepropeptides of lantibiotics differ significantly between the species, the major differences being in the region encoding part of the peptide that represent a final product prior to modifications.

The latter suggests that both structures and perhaps biological properties of these lantibiotics will be different, possibly reflecting their roles in two distinct bacterial hosts. In contrast to *pksII* and *lan1*, the *lan2* gene cluster was surrounded by nonhomologous DNA in two isolates (Fig. 2.4), and is more conserved, suggesting a relatively recent acquisition of this gene cluster by both isolates via HGT.

As mentioned in Sect. 2.3, it seems plausible that self-transmissible GLPs carrying (and/or capable of "picking up") SM biosynthetic gene clusters are responsible for their transfer and integration into a new actinomycete host. From this point of view, it is interesting to compare the existing genomic data for actinomycetes with linear and circular chromosomes. The latter is represented by *Salinispora* spp., obligatory marine actinomycete bacteria known as producers of a novel anti-cancer drug candidate, salinosporamide (Gulder and Moore 2010). Genome analyses of several *Salinispora* strains revealed that their circular chromosomes have genomic islands, where most of the SM biosynthetic gene clusters are located (Penn et al. 2009). Moreover, such islands were shown to contain multiple IS elements and transposons, which may assist in recruiting SM gene

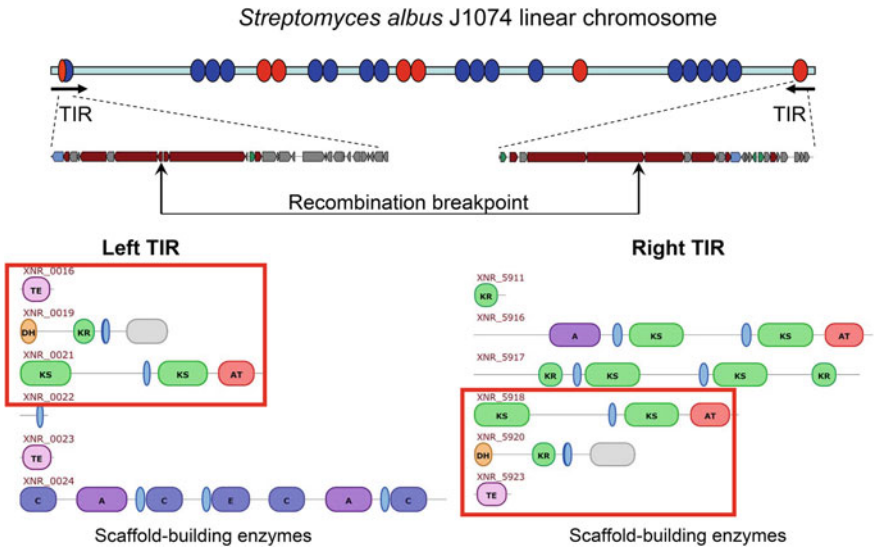


Fig. 2.5 SM gene clusters at the TIRs of *S. albus* J1074 chromosome (see text for details)

clusters to these specific locations via recombination with the incoming DNA carrying homologous sequences. In this case, a double crossover recombination would be required to stably integrate a SM gene cluster into the chromosome. Situation may be different with linear chromosomes of *Streptomyces* spp., which have terminal inverted repeats ranging in size from 35 kb in *Streptomyces davawensis* (Jankowitsch et al. 2012) to 1 Mb in *S. coelicolor* A3(2) (Weaver et al. 2004). In several *Streptomyces* spp., SM biosynthetic gene clusters are located within the TIRs, e.g., in *Streptomyces ambofaciens* (Pang et al. 2004) and *Streptomyces albus* J1074 (Zaburannyi et al. 2014; Zotchev unpublished data). This duplication of clusters at chromosomal ends may provide substrates for recombination that can fuse the existing clusters with the incoming ones carried by GLPs or those present in the vicinity of TIRs on the chromosome. This recombination may be facilitated by the presence of IS elements, which are almost always present at the TIRs of both chromosomes and GLPs (Zotchev unpublished data).

Under selective pressure from the environment, such recombination may lead to the formation of hybrid SM gene clusters that may govern biosynthesis of novel compounds beneficial to the host. Analysis of SM gene clusters at the 29-kb TIRs of a linear chromosome of *S. albus* J1074 presented in Fig. 2.5 supports the notion above. The SM gene cluster located within the right-TIR appears to be responsible for the biosynthesis of a polyketide assembled by a hexamodular PKSI enzyme complex and utilizing an amino acid as a starter. The left-TIR, however, contains genes encoding only two last modules of the same PKSI, which is now fused with what appears to be three modules of a non-ribosomal peptide synthetase (NRPS—enzymes that act in a fashion similar to PKS, but use amino acids as building blocks). An NRPS gene cluster encoding modules identical to these three modules

is also located on the chromosome close to the right-TIR, and detailed analysis allowed identification of the recombination breakpoint, where the PKS and NRPS clusters have merged.

It seems likely that this observation is not unique, and similar phenomena can be found upon analyses of other linear genomes of *Streptomyces*. Deciphering the mechanisms behind such events may not only shed light on the evolution of SM gene clusters, but also assist in rational design of novel gene clusters for biosynthesis of potential new drugs.

References

Abraham EP (1980) Fleming's discovery. Rev Infect Dis 2(1):140

Bentley SD, Brown S, Murphy LD, Harris DE, Quail MA, Parkhill J, Barrell BG, McCormick JR, Santamaria RI, Losick R, Yamasaki M, Kinashi H, Chen CW, Chandra G, Jakimowicz D, Kieser HM, Kieser T, Chater KF (2004) SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). Mol Microbiol 51(6):1615–1628

Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T (2013) antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41(Web Server issue):W204–W212

Cane DE (2010) Programming of erythromycin biosynthesis by a modular polyketide synthase. J Biol Chem 285(36):27517–27523

Challis GL, Hopwood DA (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. Proc Natl Acad Sci USA 25(100 Suppl 2):14555–14561

Chater KF (1993) Genetics of differentiation in *Streptomyces*. Annu Rev Microbiol 47:685–713

Doroghazi JR, Metcalf WW (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genom 11(14):611

Dufour YS, Donohue TJ (2012) Signal correlations in ecological niches can shape the organization and evolution of bacterial gene regulatory networks. Adv Microb Physiol 61:1–36

Flatman RH, Howells AJ, Heide L, Fiedler HP, Maxwell A (2005) Simocyclinone D8, an inhibitor of DNA gyrase with a novel mode of action. Antimicrob Agents Chemother 49(3):1093–1100

Galhardo RS, Hastings PJ, Rosenberg SM (2007) Mutation as a stress response and the regulation of evolvability. Crit Rev Biochem Mol Biol 42:399–435

Green JL, Bohannan BJ, Whitaker RJ (2008) Microbial biogeography: from taxonomy to traits. Science 320:1039–1043

Gulder TA, Moore BS (2010) Salinosporamide natural products: Potent 20S proteasome inhibitors as promising cancer chemotherapeutics. Angew Chem Int Ed Engl 49(49): 9346–9367

Hertweck C, Luzhetskyy A, Rebets Y, Bechthold A (2007) Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. Nat Prod Rep 24(1):162–190

Jankowitsch F, Schwarz J, Rückert C, Gust B, Szczepanowski R, Blom J, Pelzer S, Kalinowski J, Mack M (2012) Genome sequence of the bacterium *Streptomyces davawensis* JCM 4913 and heterologous production of the unique antibiotic roseoflavin. J Bacteriol 194(24):6818–6827

Jørgensen H, Degnes KF, Sletta H, Fjaervik E, Dikiy A, Herfindal L, Bruheim P, Klinkenberg G, Bredholt H, Nygård G, Døskeland SO, Ellingsen TE, Zotchev SB (2009) Biosynthesis of macrolactam BE-14106 involves two distinct PKS systems and amino acid processing enzymes for generation of the aminoacyl starter unit. Chem Biol 16(10):1109–1121

Jørgensen H, Degnes KF, Dikiy A, Fjaervik E, Klinkenberg G, Zotchev SB (2010) Insights into the evolution of macrolactam biosynthesis through cloning and comparative analysis of the biosynthetic gene cluster for a novel macrocyclic lactam, ML-449. Appl Environ Microbiol 76(1):283–293

Kinashi H (2011) Giant linear plasmids in *Streptomyces*: a treasure trove of antibiotic biosynthetic clusters. J Antibiot (Tokyo) 64(1):19–25

Kinashi H, Shimaji M, Sakai A (1987) Giant linear plasmids in *Streptomyces* which code for antibiotic biosynthesis genes. Nature 328(6129):454–456

Kirby R, Hopwood DA (1977) Genetic determination of methylenomycin synthesis by the SCP1 plasmid of *Streptomyces coelicolor* A3(2). J Gen Microbiol 98(1):239–252

Lawrence JG, Hendrickson H (2005) Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol 8:572–578

Mahajan GB, Balachandran L (2012) Antibacterial agents from actinomycetes—a review. Front Biosci (Elite Ed) 4:240–253

Miethke M (2013) Molecular strategies of microbial iron assimilation: from high-affinity complexes to cofactor assembly systems. Metallomics 5(1):15–28

Pang X, Aigle B, Girardet JM, Mangenot S, Pernodet JL, Decaris B, Leblond P (2004) Functional angucycline-like antibiotic gene cluster in the terminal inverted repeats of the Streptomyces ambofaciens linear chromosome. Antimicrob Agents Chemother 48:575–588

Penn K, Jensen PR (2012) Comparative genomics reveals evidence of marine adaptation in *Salinispora* species. BMC Genom 8(13):86

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10):1193–1203

Sissi C, Vazquez E, Chemello A, Mitchenall LA, Maxwell A, Palumbo M (2010) Mapping simocyclinone D8 interaction with DNA gyrase: evidence for a new binding site on GyrB. Antimicrob Agents Chemother 54(1):213–220

Subrt N, Mesak LR, Davies J (2011) Modulation of virulence gene expression by cell wall active antibiotics in *Staphylococcus aureus*. J Antimicrob Chemother 66(5):979–984

Syvanen M (2012) Evolutionary implications of horizontal gene transfer. Annu Rev Genet 46:341–358

Trefzer A, Pelzer S, Schimana J, Stockert S, Bihlmaier C, Fiedler HP, Welzel K, Vente A, Bechthold A (2002) Biosynthetic gene cluster of simocyclinone, a natural multihybrid antibiotic. Antimicrob Agents Chemother 46(5):1174–1182

Tsai HH, Huang CH, Tessmer I, Erie DA, Chen CW (2011) Linear Streptomyces plasmids form superhelical circles through interactions between their terminal proteins. Nucleic Acids Res 39(6):2165–2174

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proc Natl Acad Sci USA 104(25):10376–10381

Weaver D, Karoonuthaisiri N, Tsai HH, Huang CH, Ho ML, Gai S, Patel KG, Huang J, Cohen SN, Hopwood DA, Chen CW, Kao CM (2004) Genome plasticity in *Streptomyces*: identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. Mol Microbiol 51(6):1535–1550

Yim G, Wang HH, Davies J (2007) Antibiotics as signalling molecules. Philos Trans R Soc Lond B Biol Sci 362:1195–1200

Zaburannyi N, Rabyk M, Ostash B, Fedorenko V, Luzhetskyy A (2014) Insights into naturally minimised *Streptomyces albus* J1074 genome. BMC Genom 15(1):97

Zhou Z, Gu J, Du YL, Li YQ, Wang Y (2011) The -omics Era- toward a systems-level understanding of *Streptomyces*. Curr Genomics 12(6):404–416

Chapter 3
A Preliminary Transcriptomic Study of Galaxiid Fishes Reveals a Larval Glycoprotein Gene Under Strong Positive Selection

Graham P. Wallis and Lise J. Wallis

Abstract We describe protein sequences for a uromodulin-like larval glycoprotein (LGP) from 21 species of galaxiid fishes, with a MRCA about 30 Ma. These have been derived from both genomic DNA and cDNA, by conventional and Roche 454 sequencing. LGP shows a fast rate of evolution and an exceptionally strong signal of positive selection over the entire coding region, as evidenced by $d_N/d_S > 1$. Across all sequences, 182/336 (54 %) of residues are variable; many substitutions are profound/nonconservative and include in-frame indels. Genetic distances are, on average, 2.4x larger for coding region (996 bp) than introns (1459 bp).Our initial in situ work shows that the gene is active in developing skin and gill arches, which are likely conduits for pathogens. As ZP-domain proteins have been recently categorized as members of the immunoglobulin superfamily, we believe it is likely that LGP is an immune protein, and that these fishes are engaged in a Red Queen arms race.

3.1 Introduction

Fish of the family Galaxiidae are known for two particular features: their Gondwanan distribution and movement between rivers and the sea (diadromy) (Berra 2007; McDowall 1990). Adults live, breed, and lay their eggs in streams and rivers, but larvae are flushed out to sea when they hatch. Juvenile fish (known as whitebait) return

G. P. Wallis (✉) · L. J. Wallis
Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand
e-mail: g.wallis@otago.ac.nz

L. J. Wallis
e-mail: lise.wallis@otago.ac.nz

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_3,

to streams 4–6 months later, where they spend the rest of their lives. This migration is associated only with the larval phase, and not with breeding, and is therefore termed "amphidromy" (McDowall 1992, 2008). It is thus quite distinct from both catadromy (movement from freshwater to the sea to breed; e.g., anguillids) and anadromy (movement from the sea into freshwater to breed; e.g., salmonids). Galaxiids are an important component of biodiversity in cool-temperate southern lands, and are key players in the freshwater ecosystems of Australia and New Zealand (McDowall 1990; McDowall and Frankenberg 1981).

Although some biogeographers explained the Gondwanan distribution of galaxiids in terms of vicariance alone (Croizat et al. 1974; Rosen 1978), Darwin was more wary:

> It was formerly believed that the same fresh-water species never existed on two continents distant form each other. But Dr. Günther has lately shown that the *Galaxias attenuatus* inhabits Tasmania, New Zealand, the Falkland Islands, and the mainland of South America. This is a wonderful case, and probably indicates dispersal from an Antarctic centre during a former warm period. This case, however, is rendered in some degree less surprising by the species of this genus having the power of crossing by some unknown means considerable spaces of open ocean (Darwin 1872)

It is now clear from numerous genetic analyses that the diadromous life history of galaxiids has promoted oceanic dispersal between continents (Berra et al. 1996; Burridge et al. 2012; Wallis and Trewick 2009; Waters et al. 2000a, b). This ancestral marine phase has, however, been lost repeatedly in the group, with stream-resident species found throughout its range, in South America (McDowall 1971), South Africa (McDowall 1973), Australia (McDowall and Frankenberg 1981), New Zealand (Ling and Gleeson 2001; McDowall 1990, 2000; McDowall and Waters 2002, 2003) and New Caledonia (McDowall 1968), the only tropical representative.

Some 25 species of galaxiids live and breed in New Zealand waters, but only five of these maintain a marine juvenile phase (McDowall 2000). Completion of entire life history in streams and rivers severely limits opportunity for gene flow among adjacent river systems, and deep differentiation of these nondiadromous forms has been demonstrated (Allibone et al. 1996; Allibone and Wallis 1993; Waters et al. 2001a). Many of these stream-resident species derive from a koaro (*Galaxias brevipinnis*)-like ancestor (Burridge et al. 2012). This fact is in keeping with the propensity of koaro to climb waterfalls and penetrate deep into river systems, including glacial and volcanic lakes and alpine tarns (McDowall 1990). This trait may have promoted repeated propagation of stream-resident forms: *G. paucispondylus, G. prognathus, G. divergens, G. cobitinis, G. macronasus*, and the *G. vulgaris* group. It is now becoming clear that larvae of even the diadromous galaxiids often fail to reach the sea (Hicks et al. 2010; Hicks et al. submitted), particularly when lakes are downstream of the spawning site, so there is an inbuilt tendency to evolve freshwater lineages.

3.1.1 Population Genetic Differentiation in Diadromous Versus Stream-Resident Species

Species of fish that go to sea as juveniles have the opportunity to maintain gene flow over a wide area (Ward et al. 1994). In support of this prediction, there is little or no evidence for population genetic structuring within diadromous New Zealand galaxiids (Allibone and Wallis 1993; Barker and Lambert 1988; Waters et al. 2000a), so gene flow among river systems is large enough to overcome genetic differentiation resulting from any natal homing that might exist.

In stark contrast, our work on stream-resident, South Island endemic, *G. vulgaris*, revealed extensive genetic differentiation among catchments (Allibone and Wallis 1993), in keeping with long-term isolation in river systems. Concordant differentiation for isozymes (Allibone et al. 1996), mitochondrial DNA (Waters and Wallis 2001a, b) and morphology (McDowall 1997; McDowall and Chadderton 1999; McDowall and Wallis 1996) has led to *G. vulgaris* (sensu lato) being replaced by a complex of at least six species (McDowall 2000) and four other Evolutionarily Significant Units (Waters and Wallis 2001a; Waters and Wallis 2001b). In one case, broad sympatry of a species pair (Waters et al. 2001b) confirms species status under a biological species concept (Mayr 1942). In another case, there is long-term parapatry with only minimal hybridization (Allibone et al. 1996). In several other cases, two or more of these species are found in the same river system with little or no evidence of hybridization, although opportunities for parapatry and sympatry would have been extensive before the introduction of salmonids fragmented distributions (Crowl et al. 1992).

3.1.2 How Many Losses of Diadromy?

The resolution of a large number of closely related species prompts us to ask how they evolved. At one extreme, each species could represent an independent loss of diadromy from a koaro-like ancestor; at the other, diadromy may only have been lost once, followed by radiation in the freshwater environment. These alternatives make different phylogenetic predictions. The former scenario predicts "comb-like" tree structure of nondiadromous species branching off a koaro like-lineage, with koaro nested inside the crown of the tree, sister to the last species that it generated. The latter would show *G. brevipinnis* branching off at the base of the evolutionary tree, sister to the entire stream-resident *G. vulgaris* group.

Mitochondrial DNA analysis suggested that the answer lay somewhere in between; the radiation of the *G. vulgaris* species complex was consistent with three losses of diadromy, and thus extensive diversification in the freshwater environment alone as well (Waters and Wallis 2001a). This phylogeny was based on an extensive dataset of 5039 bp, with all nodes in the tree gaining good statistical support.

However, retention of different ancestral polymorphisms by different lineages ("lineage sorting") can lead to disagreement among gene trees (Pamilo and Nei

1988), particularly when looking at a relatively rapid radiation of numerous species. This situation can be made worse by either hybridization or selection. These three processes lead to a decoupling of gene histories both from each other, and from species histories (Ballard and Whitlock 2004). mtDNA represents a single history, so analysis of other (nuclear) genes is required to be convincing. More recent analysis of the molecular phylogenetics of the *G. vulgaris* group has included three nuclear genes (*S7*, *RAG-1*, *Numt*), and in contrast to the earlier mtDNA tree, places *G. brevipinnis* sister to the entire *G. vulgaris* group (Waters et al. 2010). That is, broader nuclear gene analysis points to a single loss of diadromy. But is this biologically plausible? Once a nondiadromous lineage has evolved in isolation, how can it spread and diversify into other freshwater systems if its young no longer migrate to sea?

3.1.3 Allopatric Speciation by Vicariant Geological Processes?

The answer may lie in the turbulent geological history of New Zealand, and of South Island in particular (Graham 2008). Its position on the Pacific and Indo-Australian tectonic plate boundary has led to extensive faulting and uplift. Through rapid uplift, erosion, or tilting of surfaces (or some combination), one river catchment can "capture" the headwater of another (Bishop 1995; Craw et al. 2003, 2008; Mortimer and Wopereis 1997). Specific hydrological scenarios can be tested by seeing whether species distributions match ancient or current connections (Mayden 1988). With DNA sequence data, one can make more fine-scale predictions and potentially apply molecular clocks to compare genetical and geological timing. We have looked at multiple locations around New Zealand where headwater capture is anticipated from geological evidence, and conclude that headwater capture in association with faunal capture has happened several times (Burridge et al. 2006, 2007, 2008; Craw et al. 2007; Waters et al. 2001a, 2006). However, there must still have been multiple losses of diadromy globally to explain the existence of stream-resident species on all major southern lands (Burridge et al. 2012), in multiple genera (Waters and McDowall 2005), and indeed on offshore islands (McDowall 2004).

3.1.4 Speciation at the Molecular Level—What Makes a Fish Stay in a Stream?

Loss of diadromy clearly leads to increased structuring and fragmentation of gene pools, culminating in speciation. Genetic changes underlying such a fundamental change in life history fulfill all the requirements of speciation genes (Nosil and Schluter 2011). That is, they are by definition causative agents of speciation rather

than its effects. But what is the underlying genetic cause of loss of diadromy? Nearly 40 years ago, it was suggested that protein sequence differences between humans and chimps were likely to be too trivial to explain the differences in anatomy and behaviour, and instead, differences in gene expression may be important (King and Wilson 1975). This perceptive and prophetic view is gaining traction as our understanding improves through the application of genomics. Adaptation and speciation may be mediated by changes in transcription factors, binding sites, promoters, repressors, enhancers, regulatory proteins, micro RNAs, and epigenetic modification rather than by changes in structural genes (Pagel and Pomiankowski 2008; van Straalen and Roelofs 2006). It seems that both structural gene changes and their regulation require consideration when searching for the molecular basis of speciation (Hoekstra and Coyne 2007).

Next generation sequencing methods have freed us to work more widely on nonmodel organisms for which microarrays might not be available (Derome and Bernatchez 2006; Vera et al. 2008). Deep sequencing of cDNA allows us to assess the abundance and sequence of mRNAs from any particular tissue of any organism, in theory giving access to any transcriptome. As loss of diadromy has occurred repeatedly in galaxiids, it was our hope that comparison of related diadromous and nondiadromous pairs of species might reveal some candidates for the genetic mechanism of this life history shift. We had access to the first 454 sequencing platform in 2007, so embarked on a project to search for speciation genes underlying the life history switch to freshwater. In the process, we serendipitously discovered a fast-evolving zona pellucida gene exhibiting extreme Darwinian selection, which we have called larval glycoprotein (LGP) (Wallis and Wallis 2011).

3.2 Materials and Methods

3.2.1 Next Generation Sequencing

3.2.1.1 Sources of Tissues

On hatching, larvae from diadromous species flow downstream and out to sea. In contrast, larvae of stream-resident species swim in surface shoals, maintaining their position in the stream. As there is such a clear behavioral difference from the time of hatching, we restricted our transcriptome analysis to freshly hatched larvae. This required collection of egg masses, and incubation under controlled conditions to control for environmental effect on gene expression. Eggs were collected from seven closely related species (Waters et al. 2010; Waters and Wallis 2001a, b) of the *Galaxias brevipinnis* group (*G. brevipinnis* NZ, *G. vulgaris, G. depressiceps, G. eldoni, G. gollumoides, G. anomalus, G.* sp D), as well as the more divergent *G. macronasus* (McDowall and Waters 2003), and *G. argenteus*. *G.* 'sp D' refers to a genetically distinct, but as yet undescribed member of the *G. brevipinnis* group (Allibone et al. 1996). Of these nine species, only

G. brevipinnis and *G. argenteus* are diadromous; the other seven stream-resident species probably arose from a *G. brevipinnis*-like ancestor (Waters et al. 2010).

Eggs from *G. brevipinnis* and *G. argenteus* were incubated on gauze in high humidity chambers at 10–12 °C; eggs from all other species were incubated fully immersed in aerated fresh water and held at 10–12 °C. All eggs were examined on a daily basis– any eggs that appeared to have fungal infections were removed and discarded. After 4–5 weeks, eggs from *G. brevipinnis* and *G. argenteus* were repeatedly immersed in spring water (10–12 °C) over a period of days. Hatchlings were harvested upon emergence and placed immediately in RNA Later (Ambion). Eggs incubating in water were left to hatch naturally (4–6 weeks). Daily checking of eggs and clearing of overnight hatchlings allowed new hatchlings to be placed in RNA Later within 2 hours of hatching. DNA was isolated from a sample of larvae to confirm species identity by cytochrome *b* sequence (Waters and Wallis 2001a). Some fresh adult tissues were also used for expression profiling, and these were similarly placed in RNA Later.

Subsequent to the transcriptome study, we have used genomic DNA isolated from our voucher specimens held in ethanol at −20 °C, going back to 1989. This was the source of DNA for 12 additional species: *Paragalaxias julianus*, *G. maculatus*, *G. fasciatus*, *G. postvectis*, *G. divergens*, *G. paucispondylus*, *G. cobitinis*, *G. prognathus*, *G. brevipinnis* Aus, *G. johnstoni*, *G. fontanus*, *Nesogalaxias neocaledonicus*, and *G. pullus*). The specimen of *G.* sp D used in Wallis and Wallis (2011) turned out to have an introgressed *G. gollumoides* LGP gene, so we have replaced that sequence with one more typical of the former taxon. As Australian and New Zealand *G. brevipinnis* are not monophyletic, we include both as two independent lineages.

3.2.1.2 cDNA Preparation

Total RNA was extracted using TriReagent and a Ribopure^{TM}Kit according to conditions specified by manufacturers (Ambion). Approximately, 40 mg of larvae (between 40 and 100 depending on the species) and between 20 and 40 mg of tissue was used for the total RNA extraction. Double-stranded cDNA was prepared from total RNA using a SMART PCR cDNA synthesis kit (Clontech). Total RNA from *G. brevipinnis* larvae was also used to prepare SMART RACE cDNA (Clontech) in order to perform 5' or 3' rapid amplification of cDNA ends.

3.2.1.3 Sequencing

Samples of cDNA from *G. brevipinnis* and *G. depressiceps* larvae were sequenced by GS FLX pyrosequencing (Roche). DNA reads were assembled into contigs using GS De Novo Assembler GS FLX Data Processing Software (Roche). Affinities of assembled contigs were determined by BLASTn and BLASTx searches of the NCBI database through PLAN (http://bioinfo.noble.org/plan/).

3.2.2 Genomic Sequencing and PCR

3.2.2.1 LGP Amplification

Primers for LGP (Wallis and Wallis 2011) were designed on the basis of 454-sequencing results for these two species, including introns by exon-primed intron-crossing (EPIC) PCR. Sequences were amplified from larval double-stranded cDNA with iTaq DNA polymerase (iNTRON Biotechnology) following manufacturer's specifications. Genomic DNA from 2 to 5 specimens of each species were amplified with Advantage 2 DNA polymerase (Clontech) following manufacturer's specifications.

3.2.2.2 Sequencing

PCR DNA products were prepared for sequencing using an Exonuclease1, Shrimp Alkaline Phosphatase presequencing kit (USB Corporation). DNA was sequenced using Big Dye^{TM}Terminator Version 3.1 (Applied Biosystems) and fragments separated on an ABI3730 DNA Analyzer.

3.2.2.3 qPCR

Total RNA samples from tissues and larvae were used in qPCR reactions. RNA was reverse-transcribed using random nonomer and oligo dT primers with Superscript III Reverse Transcriptase (Invitrogen) following manufacturer's instructions. A region spanning an intron/exon boundary was then amplified with SensiMix*Plus*SYBR Kit (Quantace) and the products analysed using a Stratagene Mx3000P Real Time PCR System.

3.2.2.4 Sequence Analysis

Sequences were aligned with CLC Sequencer Viewer 6 (www.clcbio.com) and Se-Al v2.0a11 (http://evolve.zoo.ox.ac.uk). Calculation of *d*-values and tests of selection were performed using MEGA 4 (www.megasoftware.net) (Tamura et al. 2007).

3.2.3 In Situ Hybridizations

Whole-mount in situ hybridizations were carried out on galaxiid larvae, aged 0–25 days post-hatching. RNA probes were labeled with digoxygenin-UTP by in vitro transcription with SP6 and T7 RNA polymerases. Anti-digoxygenin AP Fab fragments were used to detect the presence of labeled probe and visualized using a purple indigo dye as a byproduct from a substrate for alkaline phosphatase.

3.3 Results

3.3.1 Gene Expression Data

BLAST searches of our transcriptome sequences found matches for 50–60 % of contigs, representing types of genes that we would expect to be expressed at this early stage (developmental, mitochondrial, ribosomal, osmoregulatory, muscling, general metabolic enzymes). Initial data reveal some large differences in gene expression between the first pair of species sequenced (*G. brevipinnis*, diadromous; *G. depressiceps*, nondiadromous; Table 3.1). Some of the findings were particularly bizarre. For instance, the 124 reads for reticulon 4 (RTN4) in *G. depressiceps* covered the entire transcript (>2,400 bp), whereas the 8,014 reads in *G. brevipinnis* were confined solely to the short 3' UTR. There was also an approximately sixfold difference between species for ribosomal subunit protein 4, which seems odd. When we attempted to validate differences for 17 of these genes (performed across 4–8 species) by qPCR, none were verifiable. In fact, the only significant expression difference that we could show (again, for RTN4) was in the opposite direction to the 454 sequencing results! Namely, 454 sequencing suggested 80-fold higher expression in *brevipinnis* than *depressiceps*, but qPCR gave a lower expression level in *brevipinnis* than all seven other species. A further issue with our total larvae cDNA comparisons is that the diadromous species tend to hatch at an earlier developmental stage, when the yolk sac is much larger, so one could equally make an argument to equalize larval developmental stage between species rather than use hatch time. Expression levels of RTN4 were compared in 3-day and 7-day *G. brevipinnis* larvae to check for and age effect, but these merely echoed the findings for newly hatched larvae.

The complete absence of globin expression in *G. brevipinnis*, in contrast with abundant transcripts in *G. depressiceps* (Table 3.1) seems unlikely on the face of it, but is consistent with a developmental difference between another diadromous and nondiadromous species pair noted 45 years ago. Benzie (1968) compared developmental embryological stages of diadromous *G. maculatus* (= *G. m. attenuatus*) and nondiadromous *G. vulgaris*, and noted:

> At all stages the liver is more obvious in *G. vulgaris*, lying to the left side of the body cavity over the yolk sac. It is coloured faintly reddish by the blood corpuscles flowing through it, whereas in *G. m. attenuatus* there are so few red corpuscles that even the heart looks only a pale yellow

There were also technical issues with this early 454 platform, with ~70 % of sequences unable to be read. Taken together with the large number of differences seen and inability to validate expression level differences, we decided to take a sequence-level approach to rescue something from the data.

Table 3.1 Comparison of gene expression between larval *Galaxias brevipinnis* (103,293 reads; 5748 contigs) and *G. depressiceps* (133,436 reads; 10184 contigs) by 454 sequencing

Gene affiliation	*bre* reads/103,293	*dep* reads/133,436	*bre* : *dep* ratio
[a]Zn finger RNA binding prot	210	–	∞
[a]Tubulin polymerization-promoting prot	41	–	∞
[a]Protocadherin 19-like	667	2	431
[a]Reticulon 4	8014	124	83
[a]Sex comb on midleg-like 4	182	5	47
[a]Phosphorylase kinase β-subunit	63	3	27
[a]*Xenopus* CU075619.1 cDNA-like	649	48	17
[a]ATPase Na^+/K^+ transporting α1	333	26	17
[a]Smoothelin	191	18	14
[a]Melanoregulin	587	69	11
[a]GTP-binding prot SAR1-like	5069	620	10.6
[a]Amyloid β (A4) precursor prot	124	25	6.4
RP L4	4339	2553	2.2
[a]Sperm plasma glycoprotein (LGP)	954	582	2.1
ATPase Na^+/K^+ β	170	140	1.6
[a]Muscle CK	1221	1687	0.94
Trypsinogen	776	1327	0.76
Gapdh	418	840	0.50
[a]Mitochondrial cyt b	165	1410	0.15
Heat shock prot 90 β	205	666	0.40
dolichyl- diphospho-oligosaccharide -Protein glycosyltransferase	97	363	0.35
Distal-less homeobox 3	82	331	0.32
14 kDa apolipoprotein	142	619	0.30
ATP synthase, H transporting, Mitochondrial F1 complex, ε subunit	94	411	0.30
[a]Calmodulin 2	83	478	0.22
Annexin A1	84	486	0.22
Apolipoprotein C–I precursor	146	890	0.21
[a]Brain CK	313	2101	0.19
Mitochondrial CK	165	1410	0.15
Sideroflexin 4	3	73	0.05
α-globin	–	961	0
β-globin	–	50	0

[a] Genes were analysed by qPCR

3.3.2 Interspecific Sequence Comparisons

We first ranked the data by number of sequences that were available in each species, and looked for cases where contigs covered a large portion of the gene in both species. We then looked for contigs that showed variation among species, paying particular attention to those showing nonsynonymous substitutions. The

Table 3.2 The number of synonymous and non-synonymous differences between *G. brevipinnis* and *G. depressiceps* contigs derived from 454 sequencing of cDNA

Gene	bp	syn	non-syn	*p* (%)
Apolipoprotein B	1342	1	1	0.15
Dds glycosyltransferase	1952	3	1	0.20
Distal-less homeobox 3	836	1	1	0.24
HPS90 beta	380	1	–	0.26
Chymotrypsinogen-1 like	350	–	1	0.29
Brain CK b	1122	3	1	0.36
Mago-nashi	274	1	–	0.36
Antizyme 1	750	1	2	0.40
p21	1484	5	1	0.40
Brain CK a	1143	5	–	0.44
Photolyase	1725	3	6	0.52
Muscle CK	850	5	–	0.59
Dilute-like	507	4	–	0.79
AANAT beta	909	5	3	0.88
Aquaporin 3	620	3	4	1.13
Sperm plasma glycoprotein (LGP)	1104	4	27	2.81
[a]Mitochondrial *cox1*	998	53	2	5.51

[a] *cox1* data is given for comparison only

majority of comparisons yielded no differences, or synonymous substitutions alone, yielding a *p*-value of 0.44 % over the 15 example contigs shown (62/14,244 bp; Table 3.2). This is in (expected) contrast with mitochondrial *cox*1, with a *p*-value of 5.51 % (55/998), in keeping with the much faster rate of evolution of mtDNA (Brown 1983; Brown et al. 1979; Vawter and Brown 1986).

3.3.3 *Discovery of a Fast-Evolving Larval Glycoprotein (*LGP*)*

However, *LGP* was an outlier to this general pattern, with 27 nonsynonymous and 4 synonymous differences. Note that this is not ascribable to a higher mutation rate, since the synonymous proportion (0.362 %) is not much more than the other 15 nuclear genes (0.295 %), whereas the nonsynonymous proportion (2.44 %) is over seventeen times as big as the other 15 genes (0.14 %), and over 12 times that of mitochondrial *cox*1 (0.2 %; Table 3.2). The ratio of 27:4 is over twice that expected for a gene evolving by completely neutral evolution alone, and is strongly suggestive of positive selection. When these values are normalized so that nonsynonymous substitution rate is expressed per nonsynonymous site, and synonymous substitution per synonymous site, it enables a d_N/d_S ratio test. The test looks for significant deviations from unity (Kryazhimskiy and Plotkin 2008; Nei

and Kumar 2000), facilitating a quick screen for selection across an entire transcriptome (Bustamante et al. 2005; Fay et al. 2002; Mikkelsen et al. 2005; Plotkin et al. 2004). Purifying selection is evidenced by a value of <1, and is the most common result (Li 1997). Conversely, a value of >1 is evidence of positive selection to the extent that it causes amino acid replacement substitutions at nonsilent sites to exceed synonymous substitution at silent sites. It is thus a particularly stringent test for positive selection, especially when applied over the entire coding region of a gene.

We wanted to be certain that we were making a truly orthologous comparison, and sequenced cDNA from a further seven species for comparison, and full gene sequences from all nine species (Wallis and Wallis 2011). The main findings can be summarised as follows. *LGP* is expressed in the larvae of all nine species, and sequences were in total agreement between cDNA and genomic DNA. The gene is just over 3.1 Kbp, comprising ten exons, eight of them coding a protein of some 331–4 residues, dominated by C-terminal ZP-N plus ZP-C domains of some 261 residues, typical of zona pellucida region genes (Jovine et al. 2005). We cannot be certain about the true identity of LGP at present, but the protein is 34–39 % similar to a range of ZP-like genes, including uromodulin (Tamm-Horsfall urinary glycoprotein; *THP*) in zebrafish and seminal plasma glycoprotein (*SPP*)*120* from cichlids. It is certainly not any of the true ZP egg membrane proteins. Across all nine species, coding exons (996 bp) were evolving quicker than introns (1459 bp) by 2.4-fold. Using the modified Nei-Gojobori method (Nei and Gojobori 1986; Nei and Kumar 2000), we showed that d_N/d_S ratio was high for most exons, and significantly exceeded unity across the full coding region ($Z = 2.53$; $P = 0.006$, one-tailed).

3.3.4 *Expression and Function of* LGP

qPCR of *LGP* shows that it makes up about 0.5 % of all transcripts in larvae up to 3 weeks of age, which agrees with the 0.4–0.9 % shown by 454 sequencing (Table 3.1). Notwithstanding this abundance, it appears to be all but absent from muscle, liver, heart, brain, gill and gonad of adult *G. brevipinnis* and *G. depressiceps*. The only tissue to show any visible product was testis, but expression was about 1/500 that of larval levels. The main role of this protein seems likely to be in larvae, but we cannot rule out an additional reproductive function, for example. Reproductive genes often show elevated rates of evolution (Meslin et al. 2012; Singh et al. 2012; Swanson et al. 2003), the reasons for which are contentious (Hellberg et al. 2012). The close similarity to sperm plasma glycoprotein 120 (Gerrard and Meyer 2007) looks to be misleading with respect to function, though it could be a true ortholog having undergone specialization in cichlid fishes. *SPP120* is much larger than *LGP*, containing an upstream VWF D-domain. Furthermore, the large majority of selected sites in *SPP120* are confined to an N-terminal region upstream of *VWF*, a region not present in *LGP*.

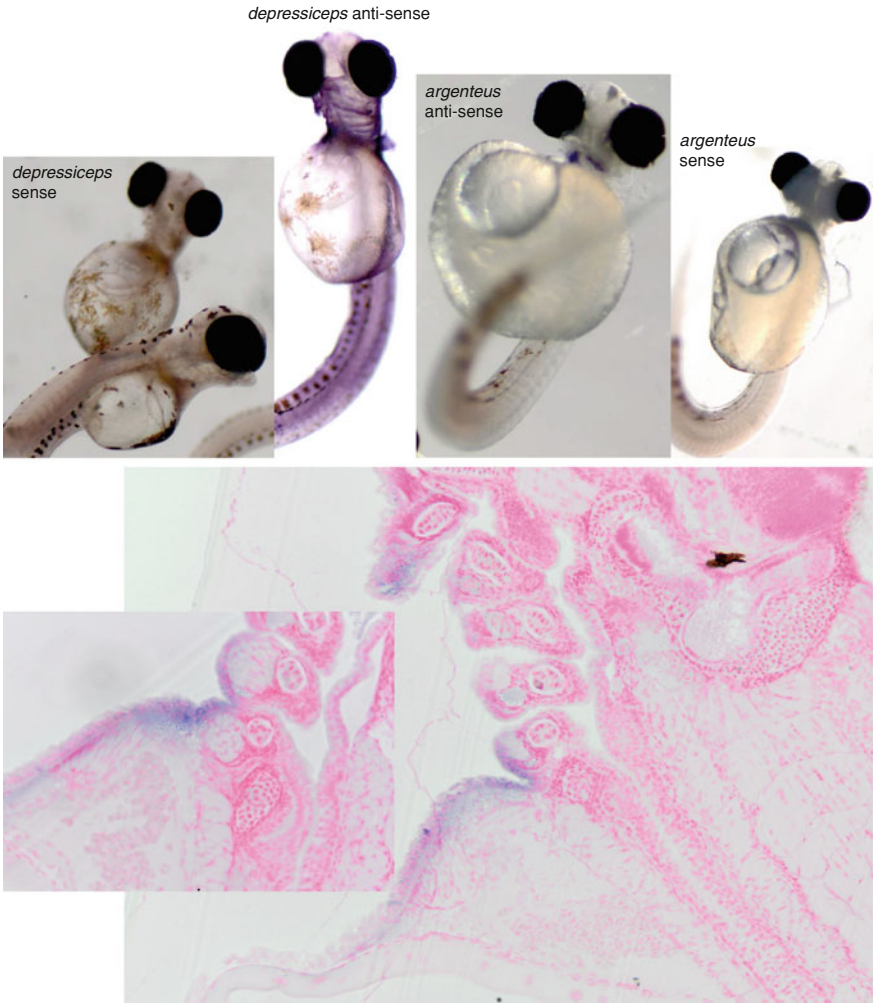


Fig. 3.1 **a** RNA in situ hybridizations for *LGP* using whole larval mounts of *Galaxias depressiceps* (non-migratory) and *G. argenteus* (migratory). Negative controls using sense strand are included. **b** Detail from *G. argenteus* in (**a**)

In situ hybridizations using newly hatched larval *G. argenteus* and *G. depressiceps* reveal widespread expression in the epidermis and developing gill arches (Fig. 3.1). These experiments are confounded by a difference in timing of development between diadromous *G. argenteus* and nondiadromous *G. depressiceps*, the former (more *r*-selected) species hatching at an earlier developmental stage. Therefore, although diadromous species eggs are less well-provisioned (Closs et al. 2013), the yolk sac is actually larger at the time of hatching.

3.3.5 Tests for Positive Selection on an Expanded Dataset

We now have 26 *LGP* sequences, representing 21 species of galaxiid, derived from genomic DNA. The fast evolution of the gene made more genetically distinct species particularly hard to amplify with the established primers, so several new ones were developed, especially in the case of *G. maculatus*. As with the first nine species examined, sequences were highly divergent (Table 3.3). Maximum composite likelihood estimation (complete deletion option) of the pattern of nucleotide substitution (Tamura et al. 2004) showed similar rates for the six classes of transversion, with two different (higher) transition rates, conforming to the Tamura-Nei model (Tamura and Nei 1993). *Galaxias maculatus* and *Paragalaxias julianus* have an exceptionally high *d*-value of 0.26, even when all indels are excluded. Congeneric comparisons involving *G. maculatus* range from 0.195 to 0.226. Even within the closely related *G. brevipinnis* group, *d*-values reach 0.041, approaching mitochondrial DNA values. At the protein level, 54 % of all amino acid positions were variable across the 21 species, which is remarkable for a clade sharing a MRCA some 30–35 Ma (Burridge et al. 2012).

As before, *Z*-tests of selection show that there is a highly significant excess of nonsynonymous substitution over the length of the gene ($Z = 2.3431$, $P = 0.0104$; pairwise deletion option). Exons 7–8 are the main contributors to this positive selection (Table 3.4A), coding the ZP-C domain, upstream of the furin cleavage site (exon 9). Exons 2–5 also show an excess of nonsynonymous substitution, in itself noteworthy, though no exon on its own shows significant excess. These values become more significant if *G. maculatus* is excluded (Table 3.4B). Exons 7–8 now show a significant signal of positive selection when analysed independently, and exons 2–6 all show a nonsignificant excess of nonsynonymous substitution. Over the entire gene, significance is close to the 0.1 % level ($Z = 3.14$, $P = 0.0011$).

To check whether selection could be attributed to any particular taxa, we additionally performed *Z*-tests for all species pairwise comparisons (Table 3.5). With the exception of *G. maculatus* (which shows exclusively negative values), the overwhelming majority of *Z*-values are positive, often significantly so; a highly unusual finding across an entire coding region.

3.4 Discussion

Our expanded dataset supports the earlier findings (Wallis and Wallis 2011), both in terms of the extent of differentiation among taxa (Table 3.3) and the widespread evidence for positive selection (Tables 3.4, 3.5). Once again, although this selection is especially attributable to exons 7–8, exons 2–5 are involved to a lesser degree (Table 3.4). In most other examples of positive selection, the effect is restricted to specific parts of the protein/gene (Hughes 1999). In the well-known case of MHC, for example, only the antigen recognition site is involved (Hughes and Nei 1988). In *SPP120*, it is almost exclusively restricted to an N-terminal region (Gerrard and

Table 3.3 Pairwise Tamura-Nei distances (*d*; lower left; × 10^3) and standard errors (upper right; × 10^3) based on 946 shared nucleotide sites of *LGP*

	jul	*macu*	*arg*	*fas*	*pos*	*div*	*macr*	*pau*	*cob*	*pro*	*bre*A	*joh*	*fon*	*bre*N	*neo*	*dep*	*vul*	*sp*D	*gol*	*ano*	*eld*	*pul*
jul	–	019	014	013	013	014	015	014	015	014	014	014	014	013	014	014	014	014	014	014	014	014
macu	262	–	016	016	016	017	017	017	018	017	017	017	017	016	017	018	017	018	017	017	017	017
arg	144	197	–	006	007	009	010	008	009	007	008	008	009	008	009	009	009	009	008	008	008	008
fas	128	195	035	–	007	009	009	008	009	007	007	008	008	007	008	008	008	008	008	008	007	008
pos	136	198	040	039	–	007	008	005	007	005	007	007	007	006	007	007	007	007	007	007	007	007
div	163	225	068	069	046	–	010	007	009	007	009	009	009	008	009	009	009	009	009	009	009	009
macr	174	223	082	076	052	078	–	008	008	008	010	010	010	009	010	010	010	010	010	010	010	010
pau	152	212	054	054	024	048	056	–	007	006	008	008	008	008	009	009	009	009	008	008	008	008
cob	170	225	076	077	047	075	050	049	–	008	010	010	010	009	010	010	010	010	010	010	010	010
pro	147	209	048	049	020	045	056	038	055	–	008	008	008	007	008	008	008	008	008	008	008	008
*bre*A	155	208	058	050	048	071	088	061	083	056	–	004	006	005	006	006	006	006	006	006	006	006
joh	155	209	061	053	049	076	089	061	085	056	017	–	006	005	006	006	006	006	006	006	005	006
fon	159	211	064	055	047	075	087	063	082	056	029	029	–	005	006	007	007	007	006	007	007	007
*bre*N	142	203	053	042	038	063	077	054	073	047	019	021	025	–	004	005	004	005	004	005	004	005
neo	154	218	064	054	046	077	090	064	086	057	032	032	036	013	–	006	005	005	006	005	005	006
dep	159	226	068	058	050	077	091	068	086	057	035	032	041	023	029	–	003	002	005	004	005	005
vul	152	220	064	054	047	074	090	064	086	056	032	029	039	018	025	011	–	003	005	003	004	005
*sp*D	156	225	065	056	048	074	089	065	084	055	032	030	041	020	027	002	009	–	005	003	004	005
gol	151	217	055	051	047	069	083	057	079	055	031	030	038	016	029	024	023	022	–	005	005	005
ano	151	216	062	053	048	072	088	062	086	056	031	030	040	019	027	012	010	010	023	–	004	004
eld	151	219	058	050	048	073	090	061	085	055	028	027	039	018	027	019	017	017	023	016	–	001
pul	151	221	061	053	050	076	089	063	084	055	030	029	041	020	029	022	019	019	025	018	002	–

Table 3.4 Results of tests for positive selection by exon for *LGP* (emboldened = significant)

A								
Exon	2	4	5	6	7	8	9	total
Z-value	0.52	1.19	0.9	−0.19	1.57	1.56	−0.14	**2.34**
P-value	0.3	0.12	0.18	1	0.06	0.06	1	**0.01**
B								
Exon	2	4	5	6	7	8	9	total
Z-value	0.54	1.19	0.58	0.59	**2.66**	**2.18**	−0.18	**3.14**
P-value	0.29	0.12	0.28	0.28	**0.004**	**0.016**	1	**0.001**
codons	12	35–37	61–64	53	83	56	28	330–334

A 23 taxa
B 22 taxa (omitting *G. maculatus*)

Meyer 2007), away from both ZP-domains. In other reproductive proteins, positive selection is often localised to binding regions of gametes (Berlin et al. 2008; Calkins et al. 2007; Swanson and Vacquier 2002; Turner and Hoekstra 2006).

As well as the entire gene being implicated (though particularly the ZP-C domain), all lineages (with the exception of *G. maculatus*) show evidence of selection. That is, positive selection is happening over the entire phylogeny extending at least 30 Ma (Burridge et al. 2012). Together with the fact that many different amino acids are used at some residues, this is symptomatic of continual evolution, rather than evolution associated with a change in gene function (Messier and Stewart 1997; Stewart et al. 1987). In the latter, one would expect an asymptote towards a more progressively more adapted state, with little or no reversal. Ours is more suggestive of a Red Queen arms race mode of evolution.

One possible reason for increased significance when *G. maculatus* is omitted that it is a divergent species, so synonymous differences start to appear while nonsynonymous substitutions start to saturate, the opposite of what might be seen in the case of purifying (i.e., negative) selection. However, *P. julianus* is also deeply differentiated, yet it does not show this effect. Additionally, all *G. maculatus* comparisons yield negative *Z*-values, despite the widespread positive selection dominating on the lineages with which it is being compared, so there appears to be definitive evidence for purifying selection, rather than merely reduced positive selection. Our working hypothesis, therefore, is that something different is happening on the *G. maculatus* lineage.

A major category of genes that repeatedly show this mode of evolution are proteins involved with defence and immunity (Alcaide and Edwards 2011; Graur and Li 2000; Hughes 1999; Nei 2005; Singh et al. 2012; Tanaka and Nei 1989). Fish have a large number of anti-microbial proteins contributing to innate immunity, particularly in the skin (Plouffe et al. 2005). The fact that our in situ hybridizations show that *LGP* is widely expressed in larval epithelium and gills is consistent with LGP being an immune protein. These fishes have a thick mucus layer as adults, which may serve as an alternative physical barrier to infection. Intriguingly, the necessity for uromodulin, a closely related ZP protein with immune activity, is because of a lack of mucus lining in the urogenital tract of

Table 3.5 Pairwise tests for positive selection among 21 species of galaxiid for *LGP*. *Z*-values (upper right) and associated *P*-values (lower left; emboldened = significant)

	jul	*macu*	*arg*	*fas*	*pos*	*div*	*macr*	*pau*	*cob*	*pro*	*bre*A	*joh*	*fon*	*bre*N	*neo*	*dep*	*vul*	*sp*D	*gol*	*ano*	*eld*	*pul*
jul	–	−2.22	0.63	1.01	1.16	0.96	0.81	0.23	0.83	1.24	1.98	1.39	1.53	1.79	2.33	1.97	1.85	1.85	1.87	0.98	1.54	1.46
macu	1	–	−1.94	−1.61	−1.1	−1.05	−0.94	−1.37	−1.27	−1.35	−0.76	−1.17	−1.29	−1.15	−0.85	−1.26	−1.27	−1.31	−1.18	−1.44	−1.47	−1.42
arg	0.27	1	–	0.77	0.61	1.02	1.09	−0.77	0.21	1.69	2.75	1.83	1.9	2.32	3.16	2.44	2.17	2.28	2.35	1.13	1.86	1.93
fas	0.16	1	0.22	–	2.43	1.88	2.01	0.92	1.53	2.94	3.56	2.5	2.23	2.42	3.55	2.91	2.33	2.74	2.29	1.51	2.32	2.41
pos	0.12	1	0.27	**0.01**	–	2.63	2.18	0.65	0.86	2.02	4.12	2.34	1.69	2.4	3.05	2.57	2.1	2.39	2.3	1.29	2.3	2.37
div	0.17	1	0.16	**0.03**	**0**	–	2.53	0.39	1.65	3.82	2.85	1.8	1.52	1.45	2.57	1.56	1.17	1.41	0.82	1.15	1.19	1.29
macr	0.21	1	0.14	**0.02**	**0.02**	**0.01**	–	1.61	1.14	1.41	3.16	2.78	1.78	1.95	2.55	2.07	1.73	1.93	1.67	1.75	1.92	1.77
pau	0.41	1	1	0.18	0.26	0.35	0.06	–	0.51	0.95	2.79	1.81	0.87	1.21	2.02	1.2	0.64	1.03	0.6	0.61	0.79	0.9
cob	0.2	1	0.42	0.06	0.2	0.05	0.13	0.31	–	0.42	2.44	1.8	1.04	1.05	1.95	1.4	1.2	1.26	0.87	0.76	1.24	1.24
pro	0.11	1	0.05	**0**	**0.02**	**0**	0.08	0.17	0.34	–	3.39	1.87	1.51	1.96	2.67	1.95	1.65	1.79	1.7	0.97	1.67	1.57
*bre*A	**0.02**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0.01**	**0**	–	0.59	2.2	1.86	3.43	1.97	1.75	1.77	1.4	0.76	1.26	1.42
joh	0.08	1	**0.03**	**0.01**	**0.01**	**0.04**	**0**	**0.04**	**0.04**	**0.03**	0.28	–	0.1	0.08	1.48	0.97	0.22	0.75	0.4	−0.64	−0.01	0.1
fon	0.06	1	**0.03**	**0.01**	0.05	0.07	**0.04**	0.19	0.15	0.07	**0.01**	0.46	–	0.14	2	1.75	1.33	1.74	0.35	0.8	1.24	1.34
*bre*N	**0.04**	1	**0.01**	**0.01**	**0.01**	0.08	**0.03**	0.11	0.15	**0.03**	**0.03**	0.47	0.44	–	2.46	2.64	1.93	2.43	1.99	0.92	1.84	1.96
neo	**0.01**	1	**0**	**0**	**0**	**0.01**	**0.01**	**0.02**	**0.03**	**0**	**0**	0.07	**0.02**	**0.01**	–	2.77	2.11	2.56	2.92	1.49	2.27	2.37
dep	**0.03**	1	**0.01**	**0**	**0.01**	0.06	**0.02**	0.12	0.08	**0.03**	**0.03**	0.17	**0.04**	**0**	**0**	–	3.67	1.5	1.77	0.91	0.79	1.28
vul	**0.03**	1	**0.02**	**0.01**	**0.02**	0.12	**0.04**	0.26	0.12	0.05	**0.04**	0.41	0.09	**0.03**	**0.02**	**0**	–	3.39	1.55	0.51	1.24	1.56
*sp*D	**0.03**	1	**0.01**	**0**	**0.01**	0.08	**0.03**	0.15	0.11	**0.04**	**0.04**	0.23	**0.04**	**0.01**	**0.01**	0.07	**0**	–	1.54	0.5	0.46	0.98
gol	**0.03**	1	**0.01**	**0.01**	**0.01**	0.21	0.05	0.28	0.19	0.05	0.08	0.34	0.36	**0.02**	**0**	**0.04**	0.06	0.06	–	0.57	0.66	1.06
ano	0.16	1	0.13	0.07	0.1	0.13	**0.04**	0.27	0.22	0.17	0.23	1	0.21	0.18	0.07	0.18	0.3	0.31	0.29	–	−0.9	−0.42
eld	0.06	1	**0.03**	**0.01**	**0.01**	0.12	**0.03**	0.21	0.11	0.05	0.11	1	0.11	**0.03**	**0.01**	0.22	0.11	0.32	0.25	1	–	1.79
pul	0.07	1	**0.03**	**0.01**	**0.01**	0.1	**0.04**	0.19	0.11	0.06	0.08	0.46	0.09	**0.03**	**0.01**	0.1	0.06	0.17	0.14	1	0.04	–

mammals (Säemann et al. 2005). Also of interest is another uromodulin-like protein ("pirica") in *Rana pirica*, which in the presence of predatory salamander larvae causes swelling of tadpoles (the "bulgy morph"), thus protecting them from ingestion by the gape-limited larvae (Mori et al. 2009). The protein is thought to function through production of a gel-like matrix produced by polymerization of ZP domains, leading to water retention.

Further circumstantial evidence at the level of the protein is that the ZP-N fold has been recognized as a new subtype of the immunoglobulin superfamily (Monné et al. 2008). The potential agents of selection could be viral, bacterial, fungal or metazoan. Several metazoan parasites are known to infect galaxiid fishes, in particular two widespread skin-penetrating trematodes that cause deformation and mortality in larvae (Poulin et al. 2012).

3.5 Future Directions

We are currently attempting to sequence further species from this group, but primers do not work reliably outside *Galaxias*. We need as many *LGP* sequences as possible to perform analysis of selection by codon across a phylogeny. Our first estimates suggest that 20–40 amino acid sites show evidence of positive selection, depending on the methods used. Paramount to our program is determination of the identity and function of LGP. While our ISH work suggests epithelial expression in larvae, further qPCR analysis using adult tissues is also required. Additionally, we would like to use immunohistochemistry to determine the location of the protein itself. Luca Jovine's group (Karolinska Institutet, Sweden) has developed a mammalian cell line expression system for LGP, which we hope to use to generate monoclonal antibodies. Also in collaboration with Luca Jovine's group at Karolinska, we are attempting to fold our LGP sequences onto structures for other ZP proteins derived by X-ray crystallography. We are also about to perform Southerns to look for orthologs in related osmeriform and salmoniform fishes. At present, we cannot find any candidates in the salmon EST library, but we hope that genomic work on Pacific and Atlantic salmons may soon provide some. Ultimately, gene knockout and laboratory selection experiments are needed to resolve the function of the gene and fitness differences among genotypes. These can be combined with study of allelic variation within and among populations in the wild in an ecological genetic framework.

3.6 General Implications for New Zealand Endemism

Now that technology has progressed far beyond the early days of NGS, we will probably return to the question with which we started: the underlying mechanism of loss of diadromy. Such genes constitute "speciation genes," since they cause genetic isolation of populations and ultimately lead to speciation. It is clear that

most New Zealand biodiversity has not been evolving independently of other austral species since its isolation from the rest of Gondwana 85 Ma, but largely derives from migrants that have crossed the Tasman (Pole 1994; Wallis and Trewick 2009; Waters and Craw 2006; Winkworth et al. 2002). The NZ biota is therefore dominated by lineages of waifs and strays that happen to be good dispersers. The limited freshwater fish fauna is a case in point, entirely deriving from species that possess a marine life history phase (McDowall 2000; Waters et al. 2000b, 2002). If migration continues, then endemism will be low; two diadromous NZ galaxiids (koaro, inanga) also occur in Australia. It is only after cessation of gene flow, resulting from the types of genetic change that we are trying to identify, that speciation of local endemics ensues. NZ biodiversity may owe much to the emergence of developmental systems and genes that reduce gene flow.

References

Alcaide M, Edwards SV (2011) Molecular evolution of the toll-like receptor multigene family in birds. Mol Biol Evol 28:1703

Allibone RM, Crowl TA, Holmes JM, King TM, McDowall RM, Townsend CR, Wallis GP (1996) Isozyme analysis of *Galaxias* species (Teleostei: Galaxiidae) from the Taieri River, South Island, New Zealand: a species complex revealed. Biol J Linn Soc 57:107

Allibone RM, Wallis GP (1993) Genetic variation and diadromy in some native New Zealand galaxiids (Teleostei: Galaxiidae). Biol J Linn Soc 50:19

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. Mol Ecol 13:729

Barker JR, Lambert DM (1988) A genetic analysis of populations of *Galaxias maculatus* from the Bay of Plenty: implications for natal river return. N Z J Mar Freshw Res 22:321

Benzie V (1968) The life history of *Galaxias vulgaris* Stokell, with a comparison with *G. maculatus attenuatus*. N Z J Mar Freshw Res 2:628

Berlin S, Qu L, Ellegren H (2008) Adaptive evolution of gamete-recognition proteins in birds. J Mol Evol 67:488

Berra TM (2007) Freshwater fish distribution. University of Chicago Press, Chicago

Berra TM, Crowley LELM, Ivantsoff W, Fuerst PA (1996) *Galaxias maculatus*: an explanation of its biogeography. Mar Freshw Res 47:845

Bishop P (1995) Drainage rearrangement by river capture, beheading and diversion. Prog Phys Geogr 19:449

Brown WM (1983) Evolution of animal mitochondrial DNA. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer, Sunderland MA, pp 62–88

Brown WM, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci U S A 76:1967

Burridge CP, Craw D, Jack DC, King TM, Waters JM (2008) Does fish ecology predict dispersal across a river drainage divide? Evolution 62:1484

Burridge CP, Craw D, Waters JM (2006) River capture, range expansion, and cladogenesis: the genetic signature of freshwater vicariance. Evolution 60:1038

Burridge CP, Craw D, Waters JM (2007) An empirical test of freshwater vicariance via river capture. Mol Ecol 16:1883

Burridge CP, McDowall RM, Craw D, Wilson MVH, Waters JM (2012) Marine dispersal as a pre-requisite for Gondwanan vicariance among elements of the galaxiid fish fauna. J Biogeogr 39:306

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153

Calkins JD, El-Hinn D, Swanson WJ (2007) Adaptive evolution in an avian reproductive protein: ZP3. J Mol Evol 65:555

Closs GP, Hicks AS, Jellyman PG (2013) Life histories of closely related amphidromous and non-migratory fish species: a trade-off between egg size and fecundity. Freshw Biol 58:1162

Craw D, Burridge C, Waters J (2007) Geological and biological evidence for drainage reorientation during uplift of alluvial basins, central Otago, New Zealand. N Z J Geol Geophys 50:367

Craw D, Burridge CP, Upton P, Rowe DL, Waters JM (2008) Evolution of biological dispersal corridors through a tectonically active mountain range in New Zealand. J Biogeogr 35:1790

Craw D, Nelson E, Koons PO (2003) Structure and topographic evolution of the Main Divide in the Landsborough-Hopkins area of the Southern Alps, New Zealand. N Z J Geol Geophys 46:553

Croizat L, Nelson G, Rosen DE (1974) Centers of origin and related concepts. Syst Zool 23:265

Crowl TA, Townsend CR, McIntosh AR (1992) The impact of introduced brown and rainbow trout on native fish: the case of Australasia. Rev Fish Biol Fish 2:217

Darwin C (1872) The origin of species by means of natural selection. J M Dent & Sons Ltd, London

Derome N, Bernatchez L (2006) The transcriptomics of ecological convergence between 2 limnetic coregonine fishes (Salmonidae). Mol Biol Evol 23:2370

Fay JC, Wyckoff GJ, Wu CI (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415:1024

Gerrard DT, Meyer A (2007) Positive selection and gene conversion in SPP120, a fertilization-related gene, during the East African cichlid fish radiation. Mol Biol Evol 24:2286

Graham IJ (2008) A continent on the move: New Zealand geoscience into the 21st century. The Geological Society of New Zealand in association with GNS Science, Wellington, p 388

Graur D, Li W-H (2000) Fundamentals of molecular evolution. Sinauer Associates Inc, Sunderland, MA

Hellberg ME, Dennis AB, Arbour-Reily P, Aagaard JE, Swanson WJ (2012) The Tegula tango: a co-evolutionary dance of interacting, positively-selected sperm and egg proteins. Evolution 66:1681

Hicks AS, Closs GP, Swearer SE (2010) Otolith microchemistry of two amphidromous galaxiids across an experimental salinity gradient: a multi-element approach for tracking diadromous migrations. J Exp Mar Biol Ecol 394:86

Hicks AS, Waters JM, David B, Norman MD, Closs GP (submitted) Retention of pelagic larvae drives population dynamics in a widespread migratory fish. Oecologia

Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995

Hughes AL (1999) Adaptive evolution of genes and genomes. Oxford Univeristy Press, New York

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167

Jovine L, Darie CC, Litscher ES, Wassarman PM (2005) Zona pellucida domain proteins. Annu Rev Biochem 74:83

King M-C, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genet 4:e1000304

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland MA

Ling N, Gleeson DM (2001) A new species of mudfish, *Neochanna* (Teleostei: Galaxiidae), from northern New Zealand. J R Soc NZ 31:385

Mayden RL (1988) Vicariance biogeography, parsimony, and evolution in North American freshwater fishes. Syst Zool 37:329
Mayr E (1942) Systematics and the origin of species. Columbia University Press, New York
McDowall RM (1968) The status of *Nesogalaxias neocaledonicus* (Weber and De Beaufort) (Pisces: Galaxiidae). Breviora Mus Comp Zool 286:1
McDowall RM (1971) The galaxiid fishes of South America. Zool J Linnean Soc 50:33
McDowall RM (1973) The status of the South African galaxiid (Pisces, Galaxiidae). Ann Cape Prov Mus (Natural History) 9:91
McDowall RM (1990) New Zealand freshwater fishes: a natural history and guide. Heinemann Reed, Auckland
McDowall RM (1992) Diadromy: origins and definitions of terminology. Copeia 1992:248
McDowall RM (1997) Two further new species of *Galaxias* (Teleostei: Galaxiidae) from the Taieri River, southern New Zealand. J R Soc NZ 27:199
McDowall RM (2000) The reed field guide to new zealand freshwater fishes. Reed Publishing, Auckland
McDowall RM (2004) The Chatham Islands endemic galaxiid: a *Neochanna* mudfish (Teleostei: Galaxidae). J R Soc NZ 2004:315
McDowall RM (2008) Diadromy, history and ecology: a question of scale. Hydrobiologia 602:5
McDowall RM, Chadderton WL (1999) *Galaxias gollumoides* (Teleostei: Galaxiidae), a new fish species from Stewart Island, with notes on other non-migratory freshwater fishes present on the island. J R Soc NZ 29:77
McDowall RM, Frankenberg RS (1981) The galaxiid fishes of Australia. Rec Aust Mus 33:443
McDowall RM, Wallis GP (1996) Description and redescription of *Galaxias* species (Teleostei: Galaxiidae) from Otago and Southland. J R Soc NZ 26:401
McDowall RM, Waters JM (2002) A new longjaw *Galaxias* species (Teleostei: Galaxiidae) from the Kauru River, North Otago, New Zealand. N Z J Zool 29:41
McDowall RM, Waters JM (2003) A new species of *Galaxias* (Teleostei: Galaxiidae) from the Mackenzie Basin, New Zealand. J R Soc NZ 33:675
Meslin C, Mugnier S, Callebaut I, Laurin M, Pascal G, Poupon A, Goudet G, Monget P (2012) Evolution of genes involved in gamete interaction: evidence for positive selection, duplications and losses in vertebrates. PLoS ONE 7:e44548
Messier W, Stewart C-B (1997) Episodic adaptive evolution of primate lysozymes. Nature 385:151
Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang S-P, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, Archidiacono N, Peer Bork P, Butler J, Chang JL, Cheng Z, Chinwalla AT, deJong P, Delehaunty KD, Fronick CC, Fulton LL, Gilad Y, Glusman G, Gnerre S, Graves TA, Hayakawa T, Hayden KE, Huang X, Ji H, Kent WJ, King M-C, KulbokasIII EJ, Lee MK, Liu G, Lopez-Otin C, Makova KD, Man O, Mardis ER, Mauceli E, Miner TL, Nash WE, Nelson JO, Pääbo S, Patterson NJ, Pohl CS, Pollard KS, Prüfer K, Puente XS, Reich D, Rocchi M, Rosenbloom K, Ruvolo M, Richter DJ, Schaffner SF, Smit AFA, Smith SM, Suyama M, Taylor J, Torrents D, Tuzun E, Varki A, Velasco G, Ventura M, Wallis JW, Wendl MC, Wilson RK, Lander ES, Waterston RH (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69
Monné M, Han L, Schwend T, Burendahl S, Jovine L (2008) Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. Nature 456:653
Mori T, Kawachi H, Imai C, Sugiyama M, Kurata Y, Kishida O, Nishimura K (2009) Identification of a novel uromodulin-like gene related to predator-induced bulgy morph in anuran tadpoles by functional microarray analysis. PLoS ONE 4:e5936
Mortimer N, Wopereis P (1997) Change in direction of the Pelorus River, Marlborough, New Zealand: evidence from composition of Quaternary gravels. N Z J Geol Geophys 40:307
Nei M (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22:2318
Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418
Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Nosil P, Schluter D (2011) The genes underlying the process of speciation. Trends Ecol Evol 26:160
Pagel M, Pomiankowski A (2008) Evolutionary genomics and proteomics. MA, Sinauer, Sunderland, p 351
Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568
Plotkin JB, Dushoof J, Fraser HB (2004) Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. Nature 428:942
Plouffe DA, Hanington PC, Walsh JG, Wilson EC, Belosevic M (2005) Comparison of select innate immune mechanisms of fish and mammals. Xenotransplantation 12:266
Pole M (1994) The New Zealand flora—entirely long-distance dispersal? J Biogeogr 21:625
Poulin R, Closs GP, Lill AWT, Hicks AS, Herrmann KK, Kelly DW (2012) Migration as an escape from parasitism in New Zealand galaxiid fishes. Oecologia 169:955
Rosen DE (1978) Vicariant patterns and historical explanation in biogeography. Syst Zool 27:159
Säemann MD, Weichhart T, Hörl WH, Zlabinger GJ (2005) Tamm-Horsfall protein: a multilayered defence molecule against urinary tract infection. Eur J Clin Investig 35:227
Singh RS, Xu J, Kulanthinal RJ (2012) Rapidly evolving genes and genetic systems. Oxford University Press, Oxford, p 288
Stewart C-B, Schilling JW, Wilson AC (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401
Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20:18
Swanson WJ, Vacquier VD (2002) Reproductive protein evolution. Annu Rev Ecol Syst 33:161
Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4. Mol Biol Evol 24:1596
Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512
Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101:11030
Tanaka T, Nei M (1989) Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol 6:447
Turner LM, Hoekstra HE (2006) Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol Biol Evol 23:1656
van Straalen NM, Roelofs D (2006) An introduction to ecological genomics. Oxford University Press, Oxford
Vawter L, Brown WM (1986) Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. Science 234:194
Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17:1636
Wallis GP, Trewick SA (2009) New Zealand phylogeography: evolution on a small continent. Mol Ecol 18:3548
Wallis LJ, Wallis GP (2011) Extreme positive selection on a new highly-expressed larval glycoprotein (LGP) gene in *Galaxias* fishes (Osmeriformes: Galaxiidae). Mol Biol Evol 28:399
Ward RD, Woodwark M, Skibinski DOF (1994) A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. J Fish Biol 44:213
Waters JM, Allibone RM, Wallis GP (2006) Geological subsidence, river capture, and cladogenesis of galaxiid fish lineages in central New Zealand. Biol J Linn Soc 88:367
Waters JM, Craw D (2006) Goodbye Gondwana? New Zealand biogeography, geology, and the problem of circularity. Syst Biol 55:351
Waters JM, Craw D, Youngson JH, Wallis GP (2001a) Genes meet geology: fish phylogeographic pattern reflects ancient, rather than modern, drainage connections. Evolution 55:1844
Waters JM, Dijkstra LH, Wallis GP (2000a) Biogeography of a southern hemisphere freshwater fish: how important is marine dispersal? Mol Ecol 9:1815

Waters JM, Esa YB, Wallis GP (2001b) Genetic and morphological evidence for reproductive isolation between sympatric populations of *Galaxias* (Teleostei: Galaxiidae) in South Island, New Zealand. Biol J Linn Soc 73:287

Waters JM, López JA, Wallis GP (2000b) Molecular phylogenetics and biogeography of galaxiid fishes (Osteichthyes: Galaxiidae): dispersal, vicariance, and the position of *Lepidogalaxias salamandroides*. Syst Biol 49:777

Waters JM, McDowall RM (2005) Phylogenetics of the Australasian mudfishes: evolution of an eel-like body plan. Mol Phylogenet Evol 37:417

Waters JM, Rowe DL, Burridge CP, Wallis GP (2010) Gene trees versus species trees: reassessing life-history evolution in a freshwater fish radiation. Syst Biol 59:504

Waters JM, Saruwatari T, Kobayashi T, Oohara I, McDowall RM, Wallis GP (2002) Phylogenetic placement of retropinnid fishes: data set incongruence can be reduced by using asymmetric character state transformation costs. Syst Biol 51:432

Waters JM, Wallis GP (2001a) Cladogenesis and loss of the marine life history phase in freshwater galaxiid fishes (Osmeriformes: Galaxiidae). Evolution 55:587

Waters JM, Wallis GP (2001b) Mitochondrial DNA phylogenetics of the *Galaxias vulgaris* complex from South Island, New Zealand: rapid radiation of a species flock. J Fish Biol 58:1166

Winkworth RC, Wagstaff SJ, Glenny D, Lockhart PJ (2002) Plant dispersal N.E.W.S. from New Zealand. Trends Ecol Evol 17:514

Chapter 4
Land Bridge Calibration of Rates of Molecular Evolution in a Widespread Rodent

J. S. Herman, J. Paupério, P. C. Alves and J. B. Searle

Abstract There is mounting evidence that rates of molecular evolution decay over recent timescales. Care is needed, therefore, to apply appropriate rates whenever molecular variation is analysed within a temporal context. Given their focus on recent events, intraspecific phylogeographic and demographic studies are particularly vulnerable to erroneous application of rates appropriate to longer periods of evolution and divergence. Rates for recent molecular evolution can be inferred directly from the DNA sequences themselves, but external geophysical events may also be used for calibration. In particular, the formation and loss of land bridges can provide an opportunity to calibrate intraspecific genealogies, estimate molecular rates and infer the absolute timing or scale of demographic changes. The Eurasian field vole *Microtus agrestis* is an exceptional system with which to examine recent demographic change and divergence in a wild mammal, because of its clear-cut pattern of molecular variation, being composed of three evolutionarily significant units (ESUs) that are reciprocally monophyletic for mitochondrial, sex-chromosome and autosomal markers. These three lineages are confined to northern Eurasia, southern Europe and western Iberia. The northern

J. S. Herman (✉)
Department of Natural Sciences, National Museums Scotland, Chambers Street, Edinburgh EH1 1JF, UK
e-mail: j.herman@nms.ac.uk

J. Paupério · P. C. Alves
CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus Agrário de Vairão, Universidade do Porto, Vairão 4485-661, Portugal
e-mail: joanapcastro@cibio.up.pt

P. C. Alves
e-mail: pcalves@fc.up.pt

J. B. Searle
Department of Ecology and Evolution, E139 Corson Hall, Cornell University, Ithaca, NY 14853-2701, USA
e-mail: jeremy.searle@cornell.edu

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_4,

ESU is in turn comprised of six parapatric mitochondrial lineages, one of them confined to northern Britain. The restricted distribution of this lineage can be associated with the Holocene land bridge connecting Britain with mainland Europe, which permits the temporal calibration of the genealogy and the association of demographic changes with specific climatic episodes. The resulting estimate of the mitochondrial protein-coding substitution rate is very high (ca. 4×10^{-7} substitutions/site/year), similar to mutation rates measured from pedigrees, i.e. contemporary evolution. The reliability of this estimate is considered.

4.1 Molecular Clock Rates and Phylogeography

Phylogeography relates molecular variation, generally in the form of DNA sequence data, to the spatial distribution of the organism in question. Time is fundamentally important to the interpretation of phylogeographic analyses as it permits the alignment of molecular data with external events, such as geophysical and climatic change, so that the biogeographical and evolutionary process can be examined. It is therefore surprising that many phylogeographic analyses treat temporal calibration in a relatively superficial manner, often simply applying an externally derived molecular clock rate to what are sometimes very complex genealogical models. This combination of sophisticated modelling and crude estimation of timing can lead to misplaced confidence in the phylogeographic interpretation.

Molecular rates are now known to decay across the species-population transition, by an order of magnitude or more, from the mutation rate inferred in pedigree studies to long-term nucleotide substitution rates measured between species (Ho et al. 2005, 2007a; Howell et al. 2003, 2008). The effect of misapplied substitution rates has previously been demonstrated and it may have misled substantial numbers of phylogoegraphic studies (Ho et al. 2008; Herman and Searle 2011). The importance of molecular rate decay, particularly to intraspecific phylogeographic and demographic studies, has led to a recent review of the decay phenomenon itself and its implications (Ho et al. 2011).

In view of the decay in molecular clock rates, it is clearly inappropriate to use deep splits derived from the fossil record (or rates derived from such splits in other taxa) to calibrate intraspecific genealogies. Even the extrapolation of rates between different intraspecific genealogies may be misguided, albeit to a lesser degree, because they may have been derived from calibration of a somewhat shallower or deeper genetic divergence. A more reliable way to incorporate the temporal dimension into a phylogeographic dataset of a particular species is to include molecular data from samples from different time periods, extracting ancient DNA (aDNA) from preserved material (e.g. Martínková et al. 2013). In the absence of aDNA sequences, a less direct method is to align specific nodes within the genealogy to external geophysical or climatic events (e.g. Burridge et al. 2008, Herman and Searle 2011).

4.2 The Genus *Microtus*

Voles of the arvicoline genus *Microtus* are small diurnal grass-eating rodents that may occur in very high numbers and form an important part of the diet of predatory mammals and birds. About 65 species have arisen in an extensive radiation that is both rapid and ongoing (Jaarola et al. 2004; Musser and Carleton 2005), making them an excellent model for the study of vertebrate speciation and population divergence.

Among these, the Eurasian field vole *Microtus agrestis* makes a particularly interesting model system, as it displays measurable divergence at both the levels of species differentiation and population variation as demonstrated by Herman and Searle (2011) and Paupério et al. (2012). Indeed, these and other earlier studies (Hellborg et al. 2005; Jaarola and Searle 2002, 2004) have used the Eurasian field vole to draw attention to the inherent difficulty in defining the point at which speciation occurs in a rapidly evolving small mammal. The wide geographic and climatic distribution of the Eurasian field vole is notable and extends from the warm temperate Mediterranean region and Atlantic conditions of Britain at one extreme, to the Arctic Circle in northern Europe and the extreme continentality of central Siberia at the other. This makes it an excellent species in which to examine the evolutionary processes that relate to glacial refugia and colonisation history in the face of variable and changing climatic conditions (Jaarola and Searle 2002; Herman and Searle 2011).

4.2.1 The Eurasian Field Vole Species Complex

The Eurasian field vole is comprised of three evolutionarily significant units (ESUs sensu Moritz 1994) that have been distinguished from maternally (mitochondrial), paternally (Y-chromosomal) and biparentally (X-chromosomal, autosomal) inherited molecular markers (Paupério et al. 2012). For example, maximum likelihood (ML) tree-building and other phylogenetic methods recover three well-supported lineages, representing the three ESUs, for the mitochondrial protein-coding cytochrome *b* gene (Fig. 4.1). The three ESUs are found in northern Eurasia, southern Europe and western Iberia, respectively, where they occupy parapatric distributions of very different size (Fig. 4.2). The northern Eurasian and southern European ESUs meet in a contact zone that probably extends from north-eastern France to the Balkans, while the southern European and western Iberian ESUs meet at a contact zone in the northern part of the Iberian Peninsula.

The ESUs are reciprocally monophyletic for the seven different loci that were examined (Paupério et al. 2012), which may be taken as evidence for their genetic isolation and potential specific status, following Yang and Rannala (2010). However, the relationship between the lineages varies among the loci. The most likely explanation for this inconsistency among loci is incomplete lineage sorting

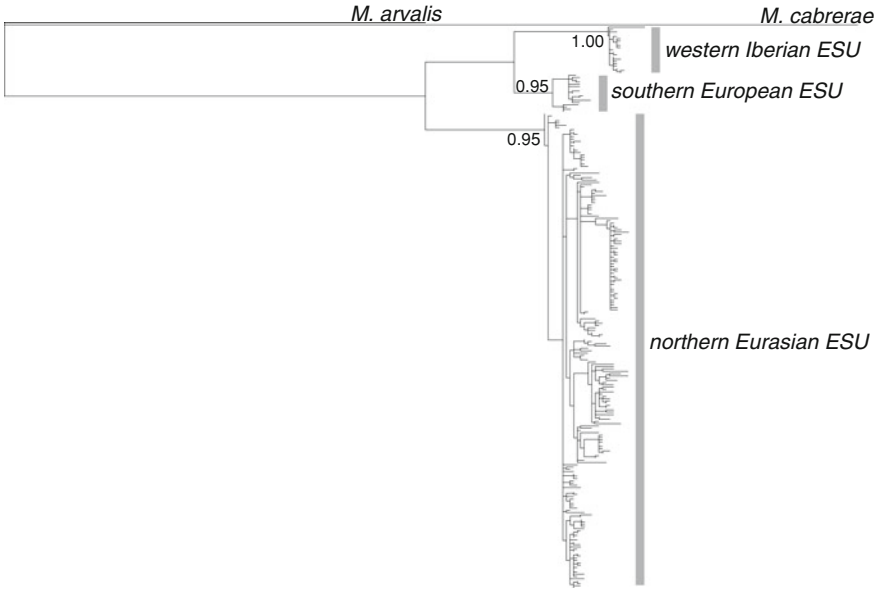


Fig. 4.1 Maximum Likelihood tree for 247 cytochrome *b* haplotypes, sampled across the range of the field vole species complex. SH-like support values for branches from approximate likelihood ratio test implemented in PhyML 3.0 (Guindon et al. 2010)

Fig. 4.2 Locations of sampled sequences assigned to the three field vole ESUs and six lineages of the northern Eurasian field vole ESU

(ILS), where the time between lineage splits has not been sufficient for ancestral polymorphism to be lost stochastically by means of genetic drift. In such circumstances, accurate estimation of species trees may be problematic, especially when dealing with recent divergence and limited genetic variation (Knowles et al. 2012). Nevertheless, the full Bayesian coalescent modelling of *BEAST (Heled

and Drummond 2010) obtained the same branching pattern among the lineages as did the comparison of ML species trees in SpedeSTEM (Ence and Carstens 2010). Interestingly, the species trees indicated that the northern Eurasian and south European lineages are closer to each other than to the western Iberian lineage, which is at odds with the mitochondrial gene tree (Fig. 4.1).

The two species tree methods differ fundamentally, in that *BEAST co-estimates the posterior distribution of species trees along with the gene trees for each of the loci, while SpedeSTEM directly compares ML trees obtained with sets of gene trees that were in turn inferred from subsampled alleles. Such consistency between methods may inspire confidence in the result (Carstens et al. 2013), but they both assume that discordance between gene trees is due to ILS and consider the lineages as separate evolutionary units rather than species per se. There is some limited evidence for hybridisation within the contact zones between the northern Eurasian and southern European ESUs (Beysard et al. 2012; Paupério et al. 2012), and between the latter and the Iberian ESU (Paupério et al. 2012). This indicates that the speciation process may have yet to run its full course, although post-mating isolation may already be well advanced (Beysard et al. 2012). For now, it seems reasonable to treat these three ESUs as distinct, if morphologically cryptic, field vole lineages with their own evolutionary trajectories.

4.2.2 The Northern Field Vole Lineage

The northern Eurasian ESU is comprised of six well-supported cytochrome *b* clades (Fig. 4.3), but this mitochondrial genetic structure has not been found in the other markers examined. This is to be expected if the divergence between these six lineages is the result of relatively recent events, given the slower rate of evolutionary change in the nuclear loci examined. As in the case of the three field vole ESUs, the cytochrome *b* lineages have geographically coherent parapatric distributions (Fig. 4.2).

The six mitochondrial lineages have similar levels of genetic diversity to one another (Table 4.1) suggesting that the lineages are of about the same age. This is borne out by comparison of the 95 % highest posterior density (HPD) range of times for the relevant nodes in the genealogy (Fig. 4.3), which were obtained with a time-rooted coalescent model.

4.3 Land Bridge Calibration

Two of the six lineages in the northern ESU, those confined to northern Britain and the southern part of the Fennoscandian Peninsula (Fig. 4.2), have restricted distributions that can be associated with land bridges of known duration. In the case of the *north Britain* lineage, there are sufficient data here for us to reliably calibrate

Fig. 4.3 Coalescent genealogy for 305 cytochrome *b* sequences, sampled across the range of the northern Eurasian field vole ESU. Maximum clade credibility tree from 760 million MCMC generations of sampling. Shaded bars are 95 % highest posterior density (HPD) ranges for times to most recent common ancestor (tMRCAs) of the six lineages. Support values are clade posterior probabilities for these

Table 4.1 Mitochondrial cytochrome *b* variation in the six clades that comprise the northern field vole ESU

Clade	n	Nucleotide diversity (π)	s.e.
Eastern	24	0.00913	0.00171
Central Europe	24	0.00747	0.00168
Scandinavia	11	0.00707	0.00142
Western	107	0.00713	0.00140
France	6	0.00697	0.00144
North Britain	133	0.00639	0.00123

Nucleotide diversity (π) with standard error (s.e.) estimated from 1,000 bootstrap pseudo-replicates (Herman and Searle 2011)

the whole cytochrome *b* genealogy, by aligning the time to the most recent common ancestor (tMRCA) of this clade with the presence of the relevant land bridge. Therefore, we focus our analysis here on the land bridge calibration associated with the *north Britain* lineage.

Much of Britain was subjected to repeated glaciations during the Pleistocene epoch and the current mammalian fauna is largely the consequence of post-glacial colonisation following the most recent and climatically severe period of the Weichselian (Devensian) stadial (Yalden 1999). The field vole is a temperate species, so the common ancestors of the *north Britain* clade would not have survived the periglacial conditions that were present in Britain and other ice-free parts of north-western Europe during the late-glacial period (Renssen and Vandenberghe 2003; Yalden 1982). The earliest date for the tMRCA of the *north Britain* clade is therefore 14,685 years BP, when the temperature rose rapidly and abruptly at the end of the Weichselian (Steffensen et al. 2008).

At this point in time, the landmass of Britain was still connected to that of mainland Europe (Yalden 1982), so field voles would have been free to colonise the British mainland as the ice sheet retreated and the climate ameliorated. However, the deglaciation also led to a rise in sea level. Its impact was exacerbated by the sinking of southern Britain, in response to the upward rebound of northern Britain after the melting of its ice cover. The land bridge was eventually lost around 8,400 years BP, with the formation of the English Channel (Lambeck 1995). This allows us to fix the most recent potential date for the tMRCA of the *north Britain* clade, as the lineage must have appeared before then, or it would form a subclade of a larger British clade.

Alignment of the *north Britain* clade tMRCA to the presence of an accessible land bridge in southern Britain, concomitant with the whole period of 6,285 years for which immigration was feasible (8,400–14,685 years BP), is sufficient to constrain the genealogy to a broad timescale. Fortunately, the calibration can be refined by placing a further constraint on the root of the genealogy, which represents the most recent common ancestor of the whole northern Eurasian field vole ESU. It seems reasonable to assume that this common ancestor would have been present before 14,685 years BP, the time of rapid warming following the onset of the final Weichselian deglaciation (Steffensen et al. 2008), as it is unlikely that any post-glacial event would reduce the effective size of the field vole population sufficiently for its coalescence.

4.3.1 Calibrated Age of Northern ESU and Clades

Using this land bridge calibration, the median tMRCA for the whole of the northern ESU was 23,629 years BP, while those of its constituent clades ranged from 10,741 to 12,324 years BP (Table 4.2). It appears that the current population of the northern field vole ESU originated during the coldest period of the last glacial maximum (LGM) and the six clades around the time of a subsequent cold

Table 4.2 Time to most recent common ancestor (tMRCA) for the whole of the northern field vole ESU (root) and its six constituent clades

Molecular rate	*M. agrestis* estimated			*Microtus* interspecific			Mammalian intergeneric		
	4.236×10^{-7} substitutions/site/year			8.333×10^{-8} substitutions/site/year			2.000×10^{-8} substitutions/site/year		
	tMRCA (years BP)			tMRCA (years BP)			tMRCA (years BP)		
Clade	Lower	Median	Upper	Lower	Median	Upper	Lower	Median	Upper
Root	16,597	**23,629**	33,918	70,241	**95,423**	127,695	291,485	**397,609**	530,290
Eastern	7,537	**11,547**	16,042	43,086	**56,348**	72,783	178,584	**234,806**	302,871
Central Europe	7,541	**11,367**	16,012	42,417	**56,740**	74,774	177,660	**236,031**	312,388
Scandinavia	6,785	**10,741**	15,059	38,529	**53,136**	70,974	160,536	**221,769**	296,753
Western	8,225	**12,324**	28,791	43,083	**58,176**	91,368	178,417	**240,385**	355,613
France	6,653	**10,874**	16,038	37,737	**54,053**	76,102	158,511	**224,234**	315,776
North Britain	9,680	**11,569**	13,444	45,577	**57,085**	71,175	189,353	**236,930**	294,788

Dates estimated by land bridge calibration or with published interspecific *Microtus* and mammalian intergeneric nucleotide substitution rates. Median and 95 % highest posterior density (HPD) limits (years BP) shown for each

period, the Younger Dryas glacial re-advance, which preceded the current Holocene warm period (Isarin 1997; Johnsen et al. 1992).

According to the demographic parameters from the coalescent simulation, there was little change in the size of the field vole population until the time of the Younger Dryas, when a rapid and sudden expansion was initiated. The distribution of the six lineages (Fig. 4.2) and the demographic pattern have together been explained by the following scenario (Herman and Searle 2011). The whole of the northern field vole ESU originated in a single bottlenecked population at the time of the LGM and was able to expand across Eurasia as the climate subsequently warmed in the Bølling-Allerød interstadial. When colder conditions returned at the Younger Dryas, the population went through a further bottleneck in six disparate locations around its new range and later expanded from these six centres at the beginning of the Holocene. The existing mitochondrial genetic structure of the field vole is therefore of entirely post-glacial origin.

4.3.2 Nucleotide Substitution Rate and Timing

The nucleotide substitution rate for the mitochondrial cytochrome *b* gene is inferred along with the other parameters of the calibrated coalescent model used here. At 4.236×10^{-7} substitutions/site/year, the rate is much higher than has typically been used in phylogeographic analyses. For example, a frequently used rate of 2×10^{-8} substitutions/site/year was derived from the divergence times, based on fossil evidence, of a range of mammals from different genera (Brown et al. 1979). An interspecific rate of 8.333×10^{-8} substitutions/site/year has been obtained from two species of *Microtus* (Triant and DeWoody 2006), but even this intrageneric rate is only one-fifth of the substitution rate inferred here. The high rate is typical of those that have been obtained by direct calibration of intraspecific genealogies that include ancient DNA sequences (Ho et al. 2007b; Martínková et al. 2013). However, the issue of rate decay remains contentious (Ho et al. 2011), and intraspecific phylogeographic studies generally continue to make use of lower interspecific rates for calibration.

To illustrate the effect of this, the MCMC simulations were repeated with a fixed molecular clock rate in lieu of the land bridge calibration. The substitution rate was fixed at either the interspecific *Microtus* estimate of 8.333×10^{-8} substitutions/site/year or the intergeneric mammalian rate of 2.0×10^{-8} substitutions/site/year. The resulting tMRCAs for both the root and clades are much earlier, especially in the latter case (Table 4.2), because the observed level of genetic divergence takes longer to accumulate when the rate of change is lower. Given that northern Eurasia was glaciated or periglacial for much of the Pleistocene (Renssen and Vandenberghe 2003), these times of origin are only plausible if some of the clades, at least the *north Britain* and *Scandinavia* ones, originated to the south of their current ranges and were replaced in their former ranges after expanding northwards. The scenario of colonisation and subsequent replacement has been

mooted previously, to explain the phylogeographic patterns of western European small mammals, including the field vole (Searle et al. 2009), however this interpretation referred to a simple sequence of population expansions after the LGM. A scenario of colonisation and subsequent replacement is difficult to reconcile with the multiple glacial and interglacial events of the Weichselian and, in the case of the widely used intergeneric rate, the Saalian and Elsterian periods (Gibbard et al. 2005). It seems unlikely that the discrete geographic distributions of the six lineages as seen would have been maintained in the face of the repeated range expansions and contractions associated with these long periods of fluctuating climate.

4.3.3 Demographic Change

The calibrated demographic change for the northern field vole ESU is modelled by the Bayesian Skyline Plot (BSP; Fig. 4.4a). There is little change in effective female population size between the origin of the whole ESU and around 14,000 years ago, at which point there is a rapid rise in population size that tails off around the middle of the Holocene. The population remains roughly static for about 5,000 years, before rising again around 3,000 years ago.

The earlier rise coincides with the sudden rise in temperature at the end of the Weichselian glacial period (Steffensen et al. 2008), which would have released large unoccupied areas of habitat in northern Europe for colonisation by the field vole, while the later rise (3,000 years ago) has been attributed to the human clearance of woodland, particularly in north-western Europe, which would favour a grassland species like the field vole (Herman and Searle 2011). While the land bridge calibrated demographic pattern is easily reconciled with climatic change, this is not the case for the alternative substitution rates (Fig. 4.4b, c). With the interspecific *Microtus* rate, the second demographic rise can be associated with the increase in temperature at the end of the Weichselian, but the earlier increase in population size is difficult to rationalise with the prevailing climate. It began about 60,000 years ago, during the coldest part of the early Weichselian, and continued through the numerous and rapid temperature shifts that followed, culminating as the temperature reached a sustained minimum just before the LGM (Johnsen et al. 2001), when the field vole would have been confined to a limited range in southern or central European refugia. Using the intergeneric mammalian rate, the two demographic expansions coincide with the onset of the Weichselian and the last glacial stage of the Saalian glacial period that preceded it (Gibbard et al. 2005). Once again, it is difficult to rationalise the growth of the field vole population with the loss of much of its potential range, as the ice sheets advanced southwards and the climate became less hospitable.

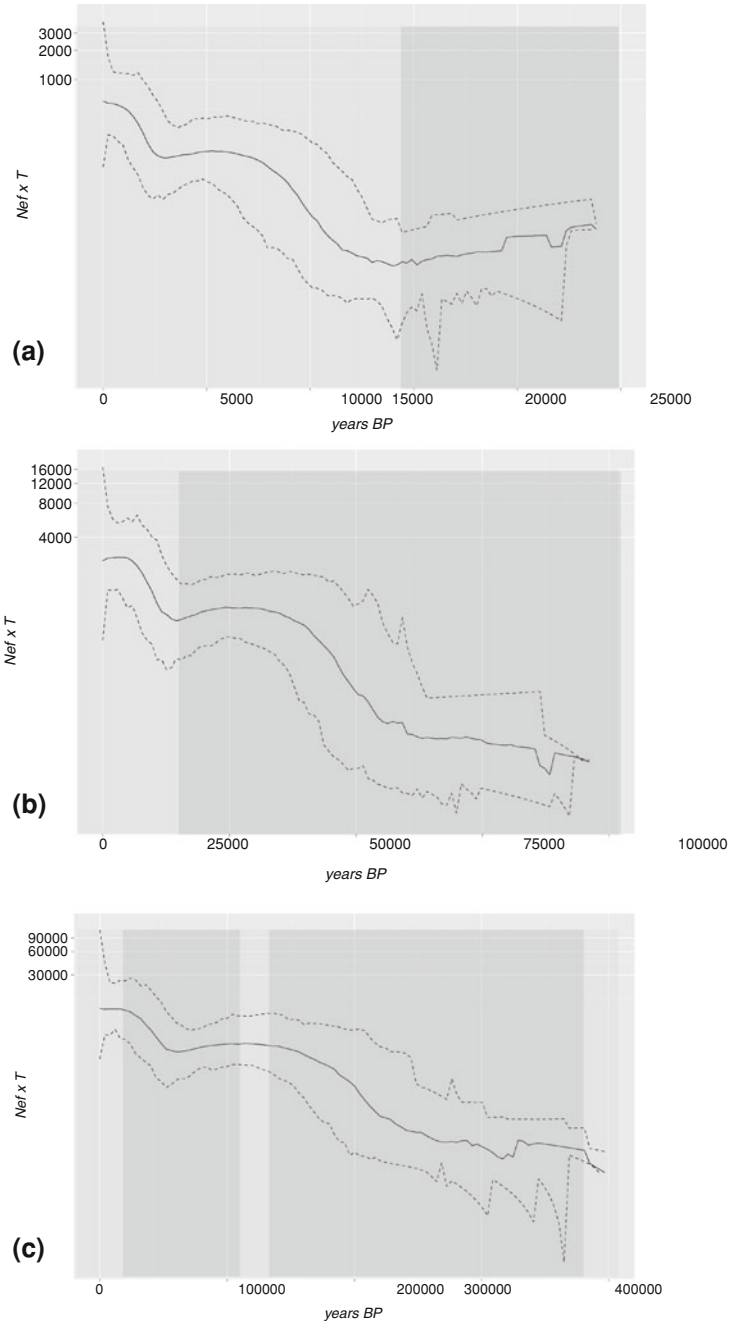


Fig. 4.4 Effective female population size of northern Eurasian field vole ESU. Bayesian skyline plots from genealogy sampling with **a.** land bridge calibration, **b.** *Microtus* interspecific and **c.** mammalian intergeneric substitution rates. Effective population size (N_{ef}), in thousands, multiplied by mean generation time (T), in years, shown on log scale for clarity. *Solid line* median. *Dashed lines upper* and *lower* 95 % HPD limits

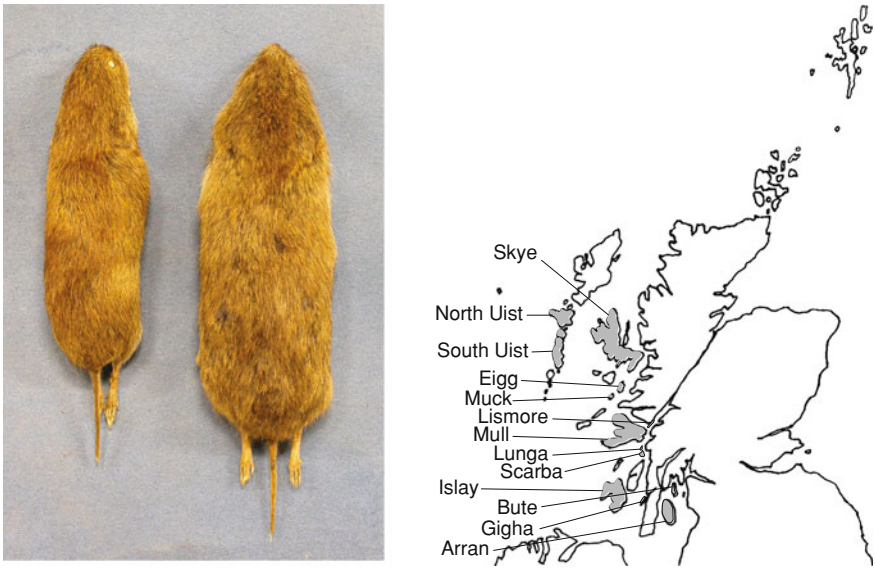


Fig. 4.5 Skins of adult female field voles from the Scottish mainland population (*left*) and Hebridean Isle of Muck (*Microtus agrestis luch* Barrett-Hamilton and Hinton). Hebridean islands with distinct cytochrome *b* haplotypes or clades. The tMRCAs for field voles colonising these islands are given in Table 4.3

4.3.4 Island Voles

Field voles from the Scottish islands, on the extreme western periphery of the field vole range, provide further evidence in support of the substitution rate that we inferred with the land bridge calibration. Small mammals from these islands have been the subject of much interest and their gigantism (Fig. 4.5), along with other unique morphological characters, led to the description of numerous supposedly distinct taxa including five named forms of the field vole (Corbet 1978). At the time of their description, the effect of the Weichselian glaciations on the islands was underestimated and their distinct morphology was attributed to a long history of isolation there. However, most if not all of the small mammals are now believed to have reached the Hebridean islands after the LGM, either by means of ice-bridges that may have been present during cold periods (White and Searle 2008) or along with animal fodder brought by human settlers in the Holocene (Corbet 1961). The pace of morphological evolution is generally considered to be faster on islands, or at least on small ones, than it is in mainland mammal populations (Millien 2006, 2011; Martínková et al. 2013) and it is therefore reasonable that the distinctive Hebridean forms of shrews, mice and voles developed over this relatively short period of time.

Distinct haplotypes and small clades from 10 of the Hebridean islands, or adjacent groups of them, are present among the cytochrome *b* sequences (Fig. 4.5). The tMRCAs obtained for these island haplogroups are shown in Table 4.3, along

Table 4.3 Time to most recent common ancestor (tMRCA) for island haplotypes and clades

Molecular rate	*M. agrestis* estimated			*Microtus* interspecific			Mammalian intergeneric		
	4.236×10^{-7} substitutions/site/year			8.333×10^{-8} substitutions/site/year			2.000×10^{-8} substitutions/site/year		
	tMRCA (years BP)			tMRCA (years BP)			tMRCA (years BP)		
Island or clade	Lower	Median	Upper	Lower	Median	Upper	Lower	Median	Upper
Uists	1,895	**3,492**	5,424	10,403	***17,139***	*26,441*	*41,122*	***70,272***	*106,891*
Skye Muck Uists	3,169	**6,298**	9,428	*16,198*	***30,051***	*45,176*	*67,900*	***126,052***	*184,095*
Muck	18	**1,238**	2,959	128	**5,531**	13,577	1,200	***21,193***	*60,509*
Eigg	72	**918**	2,228	217	**4,065**	10,601	1,094	***16,812***	*40,121*
Mull	2,263	**5,309**	9,948	11,223	***27,522***	*46,143*	*43,694*	***115,187***	*189,710*
Lismore	145	**1,145**	2,505	938	**5,358**	12,557	2,284	***21,415***	*48,716*
Scarba Lunga	133	**1,757**	3,304	1,697	**8,519**	*16,433*	2,584	***35,624***	*67,058*
Islay	655	**2,565**	4,615	5,427	**12,827**	*21,410*	*16,428*	***52,171***	*85,460*
Arran Gigha	2,223	**4,323**	7,146	12,041	***21,356***	*33,212*	*52,541*	***87,924***	*143,096*
Bute	0,229	**1,863**	3,868	1,078	**8,774**	*18,082*	6,802	***35,795***	*74,918*

Dates estimated by landbridge calibration or with published interspecific *Microtus* and mammalian intergeneric nucleotide substitution rates. Median and 95 % highest posterior density (HPD) limits (years BP) estimated from 360 to 760 million generations, combined from four independent MCMC simulations. Dates which precede the rapid post-glacial warming event (Steffensen et al. 2008) are shown in italics

with those from the simulations with the alternative interspecific and intergeneric substitution rates. The landbridge calibration determines that all of the island haplogroups originated post-glacially, consistent with the current views of their likely origin (Corbet 1961; White and Searle 2008). With the interspecific *Microtus* substitution rate, the median (and one of the lower 95 % HPD) tMRCAs for three of the island haplogroups are earlier than the deglaciation, so that would require that they originated on the mainland of Europe. For the intergeneric mammalian rate, the tMRCAs are even more problematic, with all of the median (and five of the lower 95 % HPD) tMRCAs preceding the end, and in some cases the beginning, of the Weichselian glacial period. There is no plausible mechanism by which these haplogroups might have originated in southern Europe, along with the remainder of the *north Britain* lineage, and subsequently become confined to their current restricted distributions on various islands near the west coast of Scotland.

4.4 Wider Implications of the Calibration

The land bridge calibration has shown that the phylogeographic pattern of the field vole northern ESU is post-glacial in origin, the result of genetic bottlenecks at the time of the Younger Dryas glacial re-advance (about 12,000 years ago) rather than the LGM (about 24,000 years ago). The calibration has also been used to infer the timing of events in the whole field vole species complex (Paupério et al. 2012), by means of a multilocus coalescent analysis (Heled and Drummond 2010). It appears that the field vole has become differentiated into three distinct genetic groups (the northern Eurasian, southern European and western Iberian ESUs) that are close to being separate species, all within the last 100,000 years. Small mammal species are not generally considered to have appeared within the timescale of the last (Weichselian) glaciation, so it is surprising that these lineages have become so clearly distinct, especially in the case of the northern Eurasian and southern European ESUs. However, the pattern of speciation makes sense when related to climatic events, as the lineages are likely to have originated from genetically bottlenecked populations in glacial refugia. The western Iberian lineage diverged around 70,000 years BP, which coincides with the time of maximum glacial extent in the Iberian Peninsula (Paupério et al. 2012), while the remaining lineage split around the time of the LGM, resulting in the formation of the southern European and northern Eurasian ESUs.

The field vole land bridge calibration is important because usually intraspecific genealogies are aligned with events over a longer time scale, by calibrating them with an external substitution rate that is often derived from fossils of different species. The rate estimated here (ca. 4×10^{-7} substitutions/site/year) is much higher and is similar to rates estimated from pedigrees (Howell et al. 2003) or from historical variation measured in ancient DNA samples (Ho et al. 2007b; Martínková et al. 2013). However, it was obtained by calibrating the genealogical and

demographic analysis from geophysical and climatic events, the late-glacial retreat of the ice sheet and subsequent separation of Britain from mainland Europe, which must have affected the field vole population directly. It consequently yields a result that makes intuitive sense on both a narrow (Scottish island) and a wide (Eurasian) scale.

The result here really should not be surprising, given the growing evidence for an exponential decay in molecular evolutionary rates, which is most readily explained by the effect of purifying selection on slightly deleterious mutations (Ho et al. 2008, 2011). It shows that land bridges can offer useful and direct calibrations that allow realistic estimation of intraspecific nucleotide substitution rates, phylogeographic and demographic patterns. It is interesting to speculate on the effect of such calibrations on studies of other animal species.

4.5 Methods

All sequences consisted of the whole of the 1143 base-pair protein-coding mitochondrial cytochrome *b* gene, with no ambiguous positions. The haplotypes are available on NCBI GenBank (AY167149–AY167213; FJ619746–FJ619786; GU563195–GU563299; JX284248–JX284284).

Nucleotide diversity (π), the average number of nucleotide substitutions between sequence pairs in the group, was calculated in MEGA 5.0 (Tamura et al. 2011). The Kimura 2-parameter substitution model (Kimura 1980) was used, with gamma distribution of rates across sites, and standard error was estimated from 1,000 bootstrap pseudo-replicates.

Maximum likelihood trees were inferred with PhyML 3.0 (Guindon et al. 2010). The branching pattern was robust to a range of nucleotide substitution models and the tree reported here was obtained with a GTR nucleotide substitution model (Tavaré 1986) and gamma distribution of rates across sites. SH-like support values for branches in the genealogy were based on the approximate likelihood ratio test in the PhyML 3.0 program.

Markov chain Monte Carlo (MCMC) simulations were run in BEAST 1.7–1.8 (Drummond et al. 2012). Each analysis involved four separate chains of 100 or 200 million generations, with 10 million discarded as burnin. Bayesian posterior distributions of the cytochrome *b* genealogy and other parameters in the model were obtained from the resulting 360 or 760 million samples. Logs of parameter values were examined in Tracer 1.5.0 (Rambaut and Drummond 2007), to obtain median and 95 % highest posterior density (HPD) limits. Their traces were checked for convergence and all effective samples sizes were well in excess of the 200 that is generally considered to indicate sufficient sampling. The simulations were carried out with a coalescent genealogical model, along with a 10-group skyline demographic model (Drummond et al. 2005). A two-partition (first and second codon position linked, third separate) HKY nucleotide substitution model (Hasegawa et al. 1985) was used, with gamma distribution of rates across sites, as

recommended for protein-coding genes (Shapiro et al. 2006). An uncorrelated lognormal relaxed molecular clock (Drummond et al. 2006) was used, to allow for variation in rates across branches. This was preferred over a simpler strict clock model because the 95 % HPD of the relaxed clock standard deviation did not include zero, indicating a significant departure from the strict clock model. Logcombiner and Treeannotator programs, part of the BEAST package, were used to generate Bayesian Skyline Plots and genealogies from the MCMC logs, which were examined with Tracer and FigTree 1.4 (Rambaut 2012) respectively.

For the land bridge calibration, the tMRCA of the *north Britain* clade was given a normal prior distribution with mean of 11,542 years BP and standard deviation of 1.0, truncated to upper and lower limits of 14,685 and 8,400 years BP. In addition, the root height was given a gamma distribution that peaked around 145,000 years BP, approximated with shape and scale parameters of 3.5 and 50.0 respectively, together with a lower limit of 14,685 years BP. The peak at 145,000 years BP coincides with the end of the Saalian glaciation, with which the field vole root was aligned in an earlier study (Jaarola and Searle 2002), but the result here is robust to any tested prior distribution where the root is constrained to be older than 14,685 years BP. For the alternative calibrations, the external mitochondrial substitution rates were fixed to the interspecific *Microtus* rate (8.333×10^{-8} substitutions/site/year; Triant and DeWoody 2006) and intergeneric mammalian rate (2.000×10^{-8} substitutions/site/year; Brown et al. 1979), while the tMRCA of the *north Britain* clade and the root height were not constrained to any dates.

All MCMC simulations were repeated without data, to test the combined effect of the prior parameter distributions. None of these posterior distributions were the same as those in equivalent analyses with sequence data, indicating that the results are based on the genetic data themselves. Details of all priors and input files are available on request from the authors.

References

Beysard M, Perrin N, Jaarola M, Heckel G, Vogel P (2012) Asymmetric and differential gene introgression at a contact zone between two highly divergent lineages of field voles (*Microtus agrestis*). J Evol Biol 25:400–408

Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA 76:1967–1971

Burridge CP, Craw D, Fletcher D, Waters JM (2008) Geological dates and molecular rates: fish DNA sheds light on time dependency. Mol Biol Evol 25:624–633

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. Mol Ecol 22:4369–4383

Corbet GB (1961) Origin of the British insular races of small mammals and of the 'Lusitanian' fauna. Nature 191:1037–1040

Corbet GB (1978) The mammals of the Palaearctic region: a taxonomic review. British Museum (Natural History), London

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185–1192

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973

Ence DD, Carstens BC (2010) SpedeSTEM: a rapid and accurate method for species delimitation. Mol Ecol Res 11:473–480

Gibbard PL, Smith AG, Zalasiewicz JA, Barry TL, Cantrill D, Coe AL, Cope JCW, Gale AS, Gregory FJ, Powell JH, Rawson PF, Stone P, Waters CN (2005) What status for the Quaternary? Boreas 34:1–6

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies; assessing the performance of PhyML 3.0. Syst Biol 59:307–321

Hasegawa M, Kishino H, Yano T (1985) Dating of human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27:570–580

Hellborg L, Gündüz I, Jaarola M (2005) Analysis of sex-linked sequences supports a new mammal species in Europe. Mol Ecol 14:2025–2031

Herman JS, Searle JB (2011) Post-glacial partitioning of mitochondrial genetic variation in the field vole. Proc R Soc B 278:3601–3607

Ho SYW, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22:1561–1568

Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ (2007a) Evidence for time dependency of molecular rate estimates. Syst Biol 56:515–522

Ho SYW, Kolokotronis S-O, Allaby RG (2007b) Elevated substitution rates estimated from ancient DNA sequences. Biol Lett 3:702–705

Ho SYW, Saarma U, Barnett R, Haile J, Shapiro B (2008) The effect of inappropriate calibration: three case studies in molecular ecology. PLoS ONE 3:e1615

Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A (2011) Time-dependent rates of molecular evolution. Mol Ecol 20:3087–3101

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72:659–670

Howell N, Howell C, Elson JL (2008) Time dependency of molecular rate estimates for mtDNA: this is not the time for wishful thinking. Heredity 101:107–108

Isarin RFB (1997) Permafrost distribution and temperatures in Europe during the younger Dryas. Permafrost Periglac Process 8:313–333

Jaarola M, Searle JB (2002) Phylogeography of field voles (*Microtus agrestis*) in Eurasia inferred from mitochondrial DNA sequences. Mol Ecol 11:2613–2621

Jaarola M, Searle JB (2004) A highly divergent mitochondrial DNA lineage of *Microtus agrestis* in southern Europe. Heredity 92:228–234

Jaarola M, Martínková N, Gündüz I et al (2004) Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia), inferred from mitochondrial DNA sequences. Mol Phylogenet Evol 33:647–663

Johnsen SJ, Clausen HB, Dansgaard W, Fuhrer K, Gundestrup N, Hammer CU, Iversen P, Jouzel J, Stauffer B, Steffensen JP (1992) Irregular glacial interstadials recorded in a new Greenland ice core. Nature 359:311–313

Johnsen SJ, Dahl-Jensen D, Gundestrup N, Steffensen JP, Clausen HB, Miller H, Masson-Delmotte V, Sveinbjörnsdottir AE, White J (2001) Oxygen isotope and palaeotemperature records from six Greenland ice-core stations: Camp Century, Dye-3, GRIP, GISP2, Renland and NorthGRIP. J Quaternary Sci 16:299–307

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Knowles LL, Lanier HC, Klimov PB, He Q (2012) Full modelling versus summarizing gene-tree uncertainty: Method choice and species-tree accuracy. Mol Phylogenet Evol 65:501–509
Lambeck K (1995) Late Devensian and Holocene shorelines of the British Isles and North Sea from models of glacio-hydro-isostatic rebound. J Geol Soc Lond 152:437–448
Martínková N, Barnett R, Cucchi T, Struchen R, Pascal M, Pascal M, Fischer MC, Higham T, Brace S, Ho SYW, Quéré J-P, O'Higgins P, Excoffier L, Heckel G, Hoelzel AR, Dobney KM, Searle JB (2013) Divergent evolutionary processes associated with colonization of offshore islands. Mol Ecol 22:5205–5220
Millien V (2006) Morphological evolution is accelerated among island mammals. PLoS Biol 4:e321
Millien V (2011) Mammals evolve faster on smaller islands. Evolution 65:1935–1944
Moritz C (1994) Defining 'evolutionarily significant units' for conservation. Trends Ecol Evol 9:373–375
Musser GG, Carleton MD (2005) Superfamily Muroidea. In: Wilson DE, Reeder DM (eds) Mammal species of the world. A taxonomic and geographic reference, 3rd ed. Johns Hopkins University Press, Maryland, p 834–1531
Paupério J, Herman JS, Melo-Ferreira J, Jaarola M, Alves PC, Searle JB (2012) Cryptic speciation in the field vole: a multilocus approach confirms three highly divergent lineages in Eurasia. Mol Ecol 21:6015–6032
Rambaut A (2012) FigTree v1.4. http://tree.bio.ed.ac.uk
Rambaut A, Drummond AJ (2007) Tracer v1.5. http://tree.bio.ed.ac.uk
Renssen H, Vandenberghe J (2003) Investigation of the relationship between permafrost distribution in NW Europe and extensive winter sea-ice cover in the North Atlantic ocean during the cold phases of the Last Glaciation. Quaternary Sci Rev 22:209–223
Searle JB, Kotlík P, Rambau RV, Marková S, Herman JS, McDevitt AD (2009) The Celtic fringe of Britain: insights from small mammal phylogeography. Proc R Soc B 276:4287–4294
Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23:7–9
Steffensen JP, Andersen KK, Bigler M, Clausen HB, Dahl-Jensen D, Fischer H, Goto-Azuma K, Hansson M, Johnsen SJ, Jouzel J, Masson-Delmotte V, Popp T, Rasmussen SO, Röthlisberger R, Ruth U, Stauffer B, Siggaard-Andersen M-L, Sveinbjörnsdóttir ÁE, Svensson A, White JWC (2008) High-resolution Greenland ice core data show abrupt climate change happens in few years. Science 321:680–684
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol Biol Evol 28:2731–2739
Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures Math Life Sci (Amer Math Soc) 17:57–86
Triant DA, DeWoody JA (2006) Accelerated molecular evolution in *Microtus* (Rodentia) as assessed via complete mitochondrial genome sequences. Genetica 128:95–108
White TA, Searle JB (2008) The colonization of Scottish islands by the common shrew, *Sorex araneus* (Eulipotyphla: Soricidae). Biol J Linn Soc 94:797–808
Yalden DW (1982) When did the mammal fauna of the British Isles arrive? Mammal Rev 12:1–57
Yalden DW (1999) The history of British mammals. Poyser, London
Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci USA 107:9264–9269

Chapter 5
Polyploid Speciation and Genome Evolution: Lessons from Recent Allopolyploids

Malika L. Ainouche and Jonathan F. Wendel

Abstract Whole genome duplication (polyploidy) represents a primary mechanism of sympatric speciation, especially in plants where it is particularly prevalent. In this chapter, we focus on polyploids of hybrid origin (allopolyploids) in two different genera that each have experienced recent allopolyploidy (i.e., duplication of hybrid genomes) and which have been extensively studied, namely *Spartina* (cordgrasses) and *Gossypium* (cotton genus). We review the multiple consequences of genome merger and doubling, with particular attention to temporal evolutionary changes, thus distinguishing phenomena that might characterize the earliest stages of polyploid species formation from those that are responsible for longer term changes in the contexts of natural populations (e.g., in the invasive neopolyploid *Spartina anglica*) or domestication (in the 1–2 my old allotetraploid cotton *Gossypium hirsutum*). Finally, we examine how hybridization and polyploidy enable the evolution of transgressive or novel phenotypes and ecologically relevant traits.

5.1 Introduction

Speciation is the process through which new species arise and become new independently evolving lineages as a result of reproductive isolation (Grant 1971; Mayr 1942). This process is central to the origin of biodiversity, and evolutionary

M. L. Ainouche (✉)
University of Rennes 1, UMR CNRS 6553 Ecobio, Bât. 14A,
Campus Scientifique de Beaulieu, 35042 Rennes Cedex, France
e-mail: malika.ainouche@univ-rennes1.fr

J. F. Wendel
Department of Ecology, Evolution, and Organismal Biology, Iowa State University,
Ames, IA 50011, USA
e-mail: jfw@iastate.edu

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_5,

genetic and ecological components have long been one of the most exiting research areas in biology (Dobzhansky 1937; Stebbins 1950). Speciation may be accomplished through various means, involving geographic isolation (e.g., allopatric speciation), geographically contiguous populations (parapatric speciation) or within the geographic range of the species ancestor (sympatric speciation). Reproductive isolation is attained more or less rapidly on the evolutionary time scale, through combination of extrinsic (e.g., environmental selective forces) or intrinsic (e.g., genomic events) factors. Whole genome duplication (polyploidy) represents a primary mechanism of sympatric speciation, especially in plants, where it is particularly prevalent (Cui et al. 2006; Lewis 1980).

As a consequence of polyploid speciation, more or less divergent genomes become duplicated, resulting in the coexistence of more than two chromosome sets in the same nucleus. This process often is accompanied by evolutionary novelty and transgressive phenotypic evolution, allowing the resulting polyploid species to explore new ecological niches (Pandit et al. 2011; te Beest et al. 2012; Treier et al. 2009). Polyploidy has received considerable interest in recent years, improving our understanding of the number and timing of polyploidy events and the impact of these events on genome structure and functioning (Adams and Wendel 2005; Ainouche and Jenczewski 2010; Chen and Birchler 2013; Doyle et al. 2008; Soltis and Soltis 2012).

One of the most important realizations of the genomics era is that all plants harbor in their genomes evidence of multiple ancient whole genome doubling events. An early indication that this would be the case was the discovery of the paleopolyploid nature of the first sequenced (and relatively small) "diploid" Angiosperm genome of *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000; Blanc et al. 2003). Since that time, numerous whole genome duplication (WGD) events have been documented, in different lineages at various times and phylogenetic depths (Cui et al. 2006; Jiao et al. 2011). The inescapable conclusion is that *all* plants are paleopolyploid, with polyploidy remaining an active and ongoing speciation process today.

Given this prominence of polyploidy, there has been a great deal of interest in its biological significance. Several aspects have attracted attention, including classification of the various types of polyploids, mode and frequency of formation, significance vis-à-vis adaptation and diversification, and correlations with ecological parameters (Lewis 1980; Otto and Whitton 2000; Ramsey and Schemske 1998; Stebbins 1947; Wendel and Doyle 2005). Recent work has shown that duplicated genes can contribute to functional innovation by rewiring of regulatory networks following duplication (De Smet and Van de Peer 2012), and many studies have assessed the wide spectrum of evolutionary changes affecting polyploid genomes at various evolutionary time scales, at the structural, epigenetic, transcriptional, or proteomic levels (reviewed in Chen 2007; Comai 2005; Doyle et al. 2008; Jackson and Chen 2010; Osborn et al. 2003; Wendel 2000).

The immediate and important genomic consequence of polyploidy is the simultaneous duplication of all nuclear genes. Several possible fates are acknowledged for these duplicated genes (homoeologous), including pseudogenisation,

gene loss, evolution of new functions (Ohno 1970) or subfunctionalization, where expression of the duplicated copies is partitioned between tissues or developmental stages. The gene-retention processes following WGD often are not random; for instance, the three whole-genome duplications in *Arabidopsis* have been responsible for > 90 % of the increase in transcription factors, signal transducers, and developmental genes in the last 350 million years (Maere et al. 2005). The dosage-balance hypothesis (Birchler and Veitia 2010) predicts that dosage-dependent genes are preferentially retained; more connected genes (e.g., by protein–protein interaction) are more prone to be retained following polyploidization. These various evolutionary fates of duplicate genes enhance the potential of polyploid species for functional plasticity and for exploring new evolutionary phenotypic space, so that the immediate and long-term effects of polyploidy are interconnected.

In this context, recently formed polyploids represent critical natural systems for understanding the wide spectrum of evolutionary changes set in motion by polyploidy. Accordingly, it is of interest to explore the lessons that have emerged from two very different genera that each have experienced recent allopolyploidy and which have been extensively studied, i.e., *Spartina* (cordgrasses) and *Gossypium* (cotton genus). Drawing on recent reviews we have authored (Ainouche et al. 2012; Wendel et al. 2012) we begin by reviewing our understanding of the evolutionary history of these systems and examine how hybridization and polyploidization have contributed to diversification and adaptation. We then provide a synopsis of the multiple consequences of genome merger and chromosome doubling, drawing attention to the relationships among evolutionary processes and temporal scale of divergence, and thus distinguishing phenomena that might characterize the earliest stages of polyploid species formation from those that are responsible for longer term changes. Finally, we summarize evidence that polyploidy enables the evolution of transgressive or novel phenotypes and ecologically relevant traits.

5.2 How Do Polyploid Species Form?

Polyploids may arise from genome doubling within a species ("autopolyploids") or from interspecific hybridization and genome duplication ("allopolyploidy"), which actually represent endpoints in a genetic (or taxonomic) continuum with regard to the level of divergence between the duplicated genomes (Wendel and Doyle 2005). Both autopolyploids and allopolyploids may be formed by several different mechanisms, with unreduced gamete formation at meiosis being the most frequently involved process (Ramsey and Schemske 1998). Tetraploids may form by the reunion of unreduced (diploid) gametes, or in a two-step manner involving formation of an initial triploid through the union of a reduced gamete and an unreduced gamete ("triploid bridge"). Allopolyploidy is probably more common than autopolyploidy, although the latter appears more prevalent than previously thought (Parisod et al. 2010; Ramsey and Schemske 1998). Unreduced gamete

formation is far more frequent in hybrids, which increases the likelihood of allopolyploid formation following hybridization.

5.2.1 Recurrent Polyploid Formation in Spartina

In *Spartina*, hybridization and polyploidy are particularly prominent. This grass (Poaceae, Chloridoideae) includes about 15 perennial species that have diversified mostly in the New World (Mobberley 1956). Accidental or deliberate introduction of species outside their native range over the past 150 years has accelerated diversification by facilitating hybridization with native species, introgression or speciation, resulting in several superimposed divergent genomes that coexist in the currently living species (Ainouche et al. 2012). The basic (haploid) chromosome number in this lineage is considered $x = 10$ (Marchant 1968), although all *Spartina* species recorded to date are polyploid, ranging from tetraploids to dodecaploids. Particularly fascinating is the rapid range expansion of the recently formed allododecaploid species *Spartina anglica* Hubbard, which formed in Western Europe during the end of the nineteenth century (Fig. 5.1) following hybridization between the hexaploid *Spartina alterniflora* ($2n = 62$) accidentally introduced (by ship ballast) from the Eastern America coast, with the native hexaploid *Spartina maritima* ($2n = 60$). Genome duplication in the sterile F1 hybrids *S. x townsendii* (which is still growing by vegetative means at the hybridization site) resulted in a fertile and vigorous allododecaploid species ($2n = 122–124$) (Marchant 1963) named *Spartina anglica* (Hubbard 1968), which has rapidly colonised the west-European salt-marshes and is now introduced in several continents (Ainouche et al. 2009). *Spartina anglica* is a textbook example of recent allopolyploid speciation (Aïnouche et al. 2004; Huskin 1930; Stebbins 1950). Only a few examples of similarly well-documented neopolyploidy (formed during the twentieth century) are described, including in *Tragopogon* (reviewed in Soltis and Soltis 2012), *Senecio* (reviewed in Hegarty et al. 2012), and *Cardamine* (Zozomova-Lihova et al. 2014). In natural populations, allopolyploid species may form recurrently via independent hybridization events, involving different parental genotypes and alternative maternal species, which increase the genetic diversity available to the new species ("multiple origins," sensu Soltis and Soltis 1999). For example, in the recently formed *Tragopogon* allotetraploids, at least 21 separate origins were detected for *Tragopogon miscellus* and 11 for *Tragopogon mirus* in the Palouse region (WA, USA), with reciprocal crosses involved in the origin of *T. miscellus* (Tate et al. 2009). Similarly, separate origins were documented in the British Isles for the allohexaploid *Senecio cambrensis* in North Wales and Edimburgh (Abbott and Lowe 2004). In contrast, *Spartina anglica* seems to have a single origin at Hythe, near Southampton in the UK; all populations investigated have the same maternal parent *S. alterniflora*, as evidenced from chloroplast DNA (Aïnouche et al. 2004; Baumel et al. 2001; Ferris et al. 1997), and even if multiple interspecific crosses occurred, they involved very similar genotypes as both parental species lack

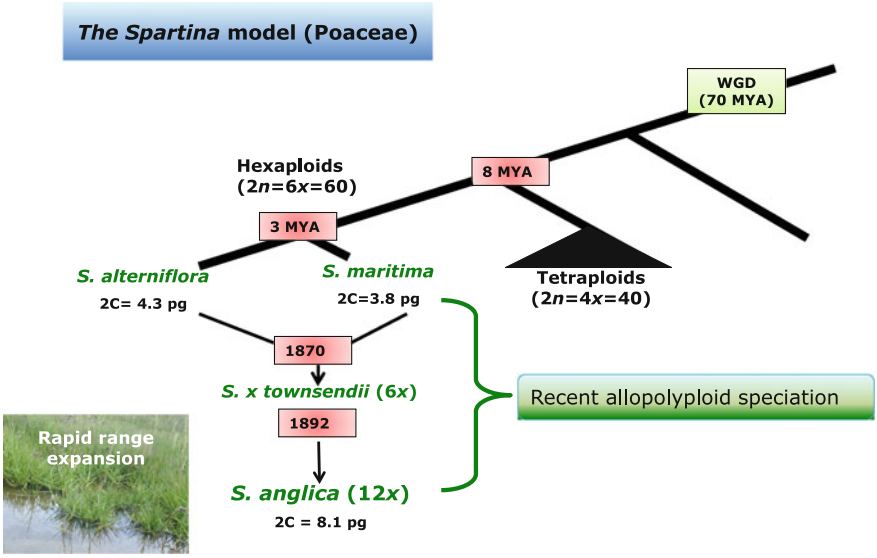


Fig. 5.1 Evolutionary history of the *Spartina* model, starting with an ancient whole genome duplication (WGD) 70 million years ago (MYA). Relevant dates are shown in the pink boxes, and genome sizes are provided in picograms (pg) for the parental species and the neo-allopolyploid

genetic diversity at the hybridization site (Baumel et al. 2001; Yannic et al. 2004). This has resulted in a strong genetic bottleneck during polyploid formation (Aïnouche et al. 2004). *Spartina anglica* most likely originated from *S. x townsendii* via unreduced gametes, although the possibility of a two-step process comparable to the triploid bridge cannot be totally ruled out (Strong and Ayres 2013), as a few nonaploid plants ($2n = 90$) with 60 chromosomes from *S. alterniflora* and 30 chromosomes from *S. maritima* were detected (Marchant 1968; Renny-Byfield et al. 2010). These nonaploid plants could also have resulted from backcrosses between the neo-dodecaploid *S. anglica* ($n = 6x$) and its hexaploid parental species *S. alterniflora* ($n = 3x$). Another independently formed, sterile F1 hybrid between *S. alterniflora* and *S. maritima* was discovered in 1892 in southwest-France in the Bidassoa Estuary (Foucaud 1897), and named *Spartina* x *neyrautii* (Jovet 1941). According to their different morphology, some authors suggested that *S. x neyrautii* and *S. x townsendii* might result from reciprocal crosses; however, molecular data revealed that both hybrids share the same chloroplast genome (Baumel et al. 2003). No genome doubling has been recorded from *S. x neyrautii*.

These recently formed F1 and allododecaploid species contain two well-differentiated genomes (Baumel et al. 2002b; Chelaifa et al. 2010b) from *S. maritima* and *S. alterniflora*. These two progenitor species diverged along the western and eastern Atlantic coasts, respectively, from a common hexaploid ancestor (Baumel et al. 2002b) during the last 3 my, as estimated from chloroplast genome sequences (Bellot et al. unpublished). The nature of polyploidy (auto versus allopolyploidy) and the origin of the hexaploid clade is not fully understood. Up to three different

duplicated (homoeologous) genes were distinguished in hexaploids for the low-copy nuclear gene *Waxy*, with substitution rates ranging from 2.18 to 4.79 % among homoeologous (Fortune et al. 2007). The presence of three different homoeologous copies of *Waxy* could support a hybrid (allopolyploid) origin of this lineage (Fortune et al. 2007). Thus, *S. anglica* contains six more or less divergent duplicated genomes, superimposed on more ancient duplications in the Chloridoideae and the Poaceae ancestor c.a. 70 mya (Paterson et al. 2012).

5.2.2 Formation of Allotetraploid Gossypium

Based on molecular data, the origin of allopolyploid *Gossypium* (the cotton genus) is suggested to have occurred in the mid-Pleistocene, perhaps 1.5 Mya (Cronn et al. 2002; Flagel et al. 2012; Senchina et al. 2003; Wendel 1989; Wendel and Cronn 2003), following a remarkable *trans*-oceanic dispersal of an Asiatic "A-genome" species ($2n = 2x = 26$) to the New World. The A-genome species hybridized with a native New World, 'D-genome' diploid ($2n = 2x = 26$), with the D-genome serving as the paternal parent and the A-genome serving as the maternal parent (Small and Wendel 1999; Wendel 1989). This resulted in an allotetraploid ($2n = 4x = 52$) lineage of AADD genomic composition (Fig. 5.2) Evidence indicates that *G. raimondii* is the closest living relative of the D genome donor, whereas the two extant A-genome species, *G. arboreum* and *G. herbaceum*, are phylogenetically sister to each other and hence equidistant from the A genome of allopolyploid cotton (Wendel and Cronn 2003). Thus, the actual parents of the allopolyploids are extinct, and reference to their parentage is more appropriately framed in terms of closest living descendants of the donor species. All analyses of DNA sequence data support a single origin of the allotetraploid lineage (Grover et al. 2012).

Following polyploidization, there was subsequent radiation into six species: *G. hirsutum* (Upland cotton), *G. barbadense* (Pima cotton, Sea Island cotton), *G. darwinii*, *G. tomentosum, G. mustelinum,* and *G. ekmanianum* (Wendel et al. 2012). *Gossypium hirsutum* is widely distributed in Central and northern South America, the Caribbean, and even reaches distant islands in the Pacific (Solomon Islands, Marquesas). Presently it is responsible for over 90 % of the cotton crop internationally, having spread from its original home in Mesoamerica to over 50 countries in both hemispheres (Brubaker et al. 1999a; Brubaker and Wendel 1994). *Gossypium barbadense* is another cultivated species, having a more southerly indigenous range centered in the northern third of South America but with a large region of range overlap with *G. hirsutum* in the Caribbean.

As mentioned above, plant evolution is characterized by repeated rounds of both shared and lineage-specific whole genome doubling. Evolutionary footprints of paleopolyploidy are also present in "diploid" ($n = 13$) cotton, as convincingly shown by the recently published genome sequence for *G. raimondii* (Paterson et al. 2012). These ancient events, superimposed on the most recent neopolyploidization leading to modern allopolyploid cotton, have profoundly impacted morphological,

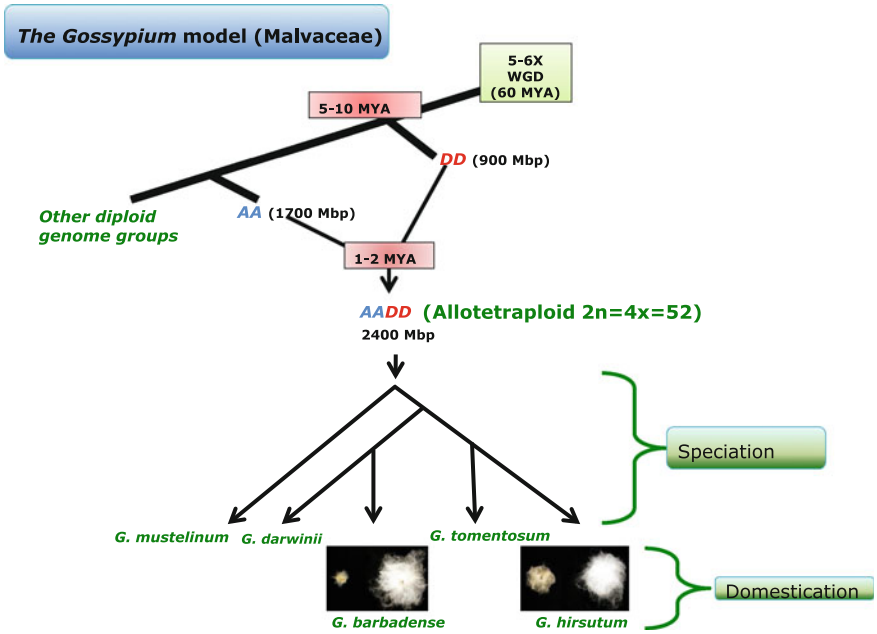


Fig. 5.2 Evolutionary history of the *Gossypium* model, starting with an ancient whole genome multiplication (WGD) 60 million years ago (MYA). Relevant dates are shown in the pink boxes, and genome sizes are provided in Megabase pairs (Mbp) for the diploid ancestors and the allotetraploids. Following divergence at the allopolyploid level, two species were independently domesticated for their seed trichomes (cotton), with representative phenotypes of wild and domesticated populations shown

ecological and physiological diversification of the genus. More recently, humans independently domesticated four species, including two allotetraploids from the Americas, *G. hirsutum* and *G. barbadense* (Brubaker et al. 1999a; Hutchinson 1951, 1954; Hutchinson et al. 1947; Percy and Wendel 1990; Wendel et al. 2009). This phylogenetic and temporal perspective provides the organismal framework for analyses of the consequences of polyploid evolution, including those focused on genomic, epigenomic, and phenotypic levels.

5.3 Genetic and Genomic Consequences of Allopolyploid Speciation

5.3.1 Structural Genome Evolution

In contrast to several young or experimentally re-synthesized allopolyploid systems that exhibit rapid structural evolution (Ozkan et al. 2001; but see Arnaud et al. 2013; Gaeta et al. 2007; Lim et al. 2008; Skalicka et al. 2005; Szadkowski et al. 2010;

Tate et al. 2009), the new allo-dodecaploid *S. anglica* exhibits relative genome stability compared to its parental species (Aïnouche et al. 2004; Baumel et al. 2002a; Salmon et al. 2005). Populations of *S. anglica* are composed of the same additive multilocus genotype (with respect to its parents *S. maritima* and *S. alterniflora*) that is rapidly expanding around the world (Baumel et al. 2001). Similarly, no major structural change was detected in synthetic allotetraploid cotton, as indicated from parental fragment additivity for 22,000 AFLP genomic loci (Liu et al. 2001b). Comparisons of gene order and synteny using comparative mapping also demonstrate that relatively few rearrangements have arisen in the 1–2 my since allopolyploid formation (Brubaker et al. 1999b; Rong et al. 2004). Thus, the polyploid genomes of *Spartina* and *Gossypium* appear to be relatively quiescent, at least with respect to the phenomenon of rapid genome change. Genome restructuring may result from irregular chromosome pairing (including homoeologous recombination) at meiosis. *Tragopogon* neopolyploids (Chester et al. 2012; Lim et al. 2008; Tate et al. 2009) and synthetic allotetraploid *Brassica napus* (Szadkowski et al. 2010) appear particularly dynamic in this respect. Supporting the apparent structural stability in *Spartina* and *Gossypium*, regular bivalent pairing is observed in both allopolyploids (Endrizzi et al. 1985; Marchant 1968).

Another potential source of rapid restructuring may be transposable element (TE) activity. As allopolyploidization entails the merger of two different complements of TEs, it has been hypothesized that this "genomic shock" (McClintock 1984) creates the potential for TE activation due to the generalized disruption of epigenetic suppression of TE activity following the merger of two diverged regulatory systems. This might be accentuated when the parental species differ either quantitatively (e.g., different genome size and TE copy numbers) or qualitatively (different composition in TE families). In *Spartina*, the hexaploid parental species *S. maritima* ($2n = 60$) and *S. alterniflora* ($2n = 62$) have a slightly different genome size (averaging 3.8 pg and 4.3 pg respectively, Fortune et al. 2008) consistent with their respective chromosome number. The relative proportion of repetitive sequences was recently estimated as about 30.5 % of the *S. maritima* genome from 26.7 MB BAC-end sequences, where class I Gypsy-like LTR retrotransposons were predominant (Ferreira de Carvalho et al. 2013a). The relative proportion of repetitive sequences in the genome of *S. alterniflora* is not known, but some differential insertion patterns between *S. maritima* and *S. alterniflora* were detected using a transposon display method (Parisod et al. 2009). The allododecaploid *S. anglica* exhibits an additive genome size with respect to its parental species (Renny-Byfield et al. 2010), as well as additive transposon display patterns (Baumel et al. 2002a; Parisod et al. 2009). Interestingly, this absence of transposition burst following allopolyploidy is accompanied by consistent epigenetic (DNA methylation) changes detected in regions flanking TEs following hybridization (in *S. x townsendii*), that were transmitted to the neopolyploid *S. anglica* (Parisod et al. 2009). In *Gossypium*, genome sizes vary widely among diploid cotton species, from ~900 Mb in the D-genome diploids to ~2,600 Mb in the Australian diploids (Hendrix and Stewart 2005), reflecting primarily the differential and punctuated proliferation of various families

of *copia* and *gypsy* TEs, as well as lineage-specific differences in the rate of deletions (Hawkins et al. 2006, 2008, 2009). The two progenitor genomes of allopolyploid cotton differ twofold in size, and they differ in their complement of resident TEs. Genome sizes of the allotetraploid species represent about 5–10 % less than the sum of the genome sizes of present-day diploids (Fig. 5.2), likely illustrating the general phenomenon of genomic downsizing (Leitch and Bennett 2004). This observation is supported by comparative BAC sequencing analyses in allotetraploid *G. hirsutum* (Grover et al. 2004, 2007), where small deletions were noted to be more prevalent in the polyploid genomes (A_T and D_T) than in either diploid genome. However, like in *Spartina*, hybridization and polyploidy do not appear to have stimulated a massive TE proliferation in *Gossypium*. (Hu et al. 2010) used phylogenetic and quantitative methods to identify changes in TE populations in allopolyploid cotton, and showed that the major LTR retrotransposons phylogenetically clustered with either their A- or D- genome antecedent elements in a genome-specific fashion, with no evidence of an impressive, recent, TE burst.

This conclusion appears to be a general trend from the studies reported to date in recent or synthetic allopolyploids (Parisod et al. 2010). Notwithstanding the relative TE quiescence indicated by these studies, evidence using FISH (Hanson et al. 1998, 1999) implicates at least a modest level of TE activity in allopolyploid cotton. These data show that a family of *copia*-like retrotransposable elements "horizontally" transferred across genomes following allopolyploid formation, raising the possibility that this process has played a role in diversification and adaptation via novel TE insertions.

5.3.2 *Evolution (Retention–Loss) of Duplicated Homoeologous Genes*

At the DNA level, the various fates of duplicated gene may be modeled as shown in Fig. 5.3. As shown, homoeologs may evolve independently and at the same rate (the null hypotheses), or alternatively, they may interact via recombination, be lost through silencing or deletion, or evolve at different rates, for example when there is strong, homoeolog-specific directional selection.

The *c.a*. 150-years old allopolyploid *S. anglica* exhibits parental additivity for the nuclear genes analyzed so far (reviewed in Ainouche et al. 2012); however in the parental hexaploid species *S. maritima* and *S. alterniflora*, differential homoeolog retention or loss may be encountered. For instance, Fortune et al. (2007) found that *S. alterniflora* retained three divergent duplicated homoeologs (as expected from its ploidy level) at the *Waxy*—B locus, whereas only one copy was retained in *S. maritima* for the same locus. Current investigations (Salmon et al. unpublished) using high-throughput sequencing (454 Roche and Illumina High Seq) will offer more generalized view of the level of gene retention at various evolutionary time scales in this system (i.e., in the nascent dodecaploid *S. anglica* and its older hexaploid parents).

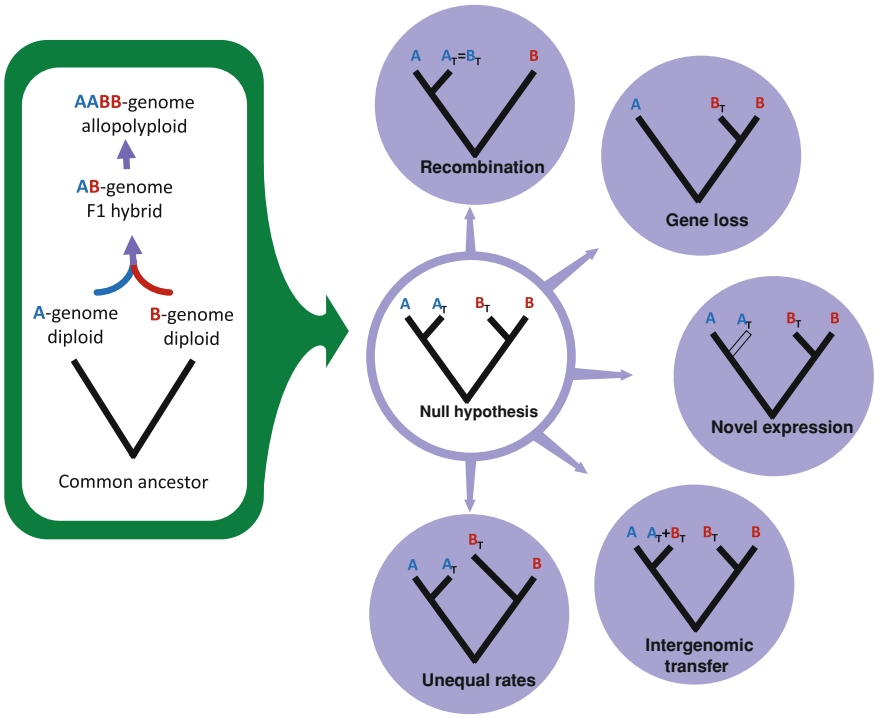


Fig. 5.3 A model of various possibilities for duplicate gene evolution following allotetraploid speciation (modified from Wendel et al. 2012) involving two diploid parental species A and B. The null expectation (*center*) derives from the organismal history (*left*): if homoeologs evolve independently and at equal rates following allopolyploid formation, then each should be phylogenetically sister to its ortholog from the donor diploid, rather than to the other homoeolog. Similarly, if rates of sequence evolution are similar at the diploid and allopolyploid level, branch lengths for the two A-genome sequences (one from the diploid, "A," and the other from the allopolyploid, "A_T") should be similar, as they should for the two B-genome sequences ("B" and "B_T"). The utility of this null hypothesis lies in its falsification; if homoeologous sequences interact via concerted evolutionary forces or nonreciprocal homoeologous recombination, for example, a different tree may be recovered ("Recombination," *top center*), or if there is strong directional selection or pseudogenization, rate inequalities may become evident ("Unequal rates," *bottom center*). Additional possibilities include loss of one of homeologs ("Gene loss," *top right*), replicative transfer of sequences from one genome to the other ("Intergenomic transfer," *bottom right*), and evolutionary divergence in duplicate gene expression domains or amounts ("Novel expression," *right middle*). This latter category, novel expression, encompasses multiple phenomenon, including developmentally or environmentally regulated biases in homoeolog expression ratios, organ or tissue or cell-specific homoeolog silencing, novel expression domains, and transgressive (higher or lower than either progenitor diploid) expression amounts

In the natural (1–2 my old) allotetraploid *Gossypium* species, early work (Cronn et al. 1999; Senchina et al. 2003) demonstrated that duplicated genes typically are both retained, and that they evolve essentially as modeled in the central panel of

Fig. 5.3, i.e., equally, additively, and at equivalent rates. A more recent analysis (Flagel et al. 2011), based on a global assembly of 5 million Sanger and 454 ESTs supplemented by ~150 million 82 bp Illumina reads, has provided an expanded view of genic evolution in diploid and allopolyploid cotton. These data, representing analysis of ~10,000 genes in each comparison, show that allopolyploidy has not, in general, been accompanied by an enhanced rate of nucleotide substitution in coding regions, as might be expected from an assumption of rapid decay of "redundant" duplicated copies. However, for some genes (Small and Wendel 1999, 2002) homoeologs may acumulate synonymous substitutions at different rates. At present, little information exists for *Gossypium* that enables a thorough analysis of the relative rates of pseudogenization or rates of gene loss in cotton, though evidence to date based on comparative BAC sequencing (Grover et al. 2004, 2007), comparative mapping analyses (Brubaker et al. 1999b; Rong et al. 2004), EST collections (Flagel et al. 2011), and AFLP studies (Liu et al. 2001a) suggests that rates of gene loss are neither high nor particularly biased with respect to genomic origin. A recent study using Southern hybridization, however, detected three losses (of 27 genes studied) of D-homoeologs from allopolyploid cotton and no losses of the A-homoeolog, suggesting a possible bias (Rong et al. 2010). Key data from ongoing whole genome sequencing likely will be generated soon that will permit these speculations to be evaluated.

5.3.3 Homoeolog Interaction

The study of Wendel et al. (1995) demonstrated interaction among the 18S–26S ribosomal genes that exist at multiple loci in the A- and D-genomes. Instead of evolving independently, as expected if homoeologous repeats did not interact, repeats at the different arrays in allopolyploid cotton have been "homogenized" to the same sequence (either "A-like" or "D-like") by one or more processes of concerted evolution (reviewed by Elder and Turner 1995). In four of the five allopolyploid species known at that time, interlocus homogenization has created exclusively D-genome like rDNAs, whereas in *G. mustelinum* nearly all rDNA repeats homogenized to an A-like form. This example showed that since polyploid formation 1–2 mya, some 3,800 repeats, each approximately 10 kb in length, were "overwritten" with the alternative form originating from the other parental genome, probably through unequal crossing over or gene conversion, and that this phenomenon operated bi-directionally, in different directions in different allopolyploid lineages. This homogenization of rDNA genes has since been well-documented in many other allopolyploids, including those in *Nicotiana* (Kovarik et al. 2004) and *Tragopogon* (Soltis et al. 2004). In the *Spartina* F1 hybrid (*S. x townsendii*) and the allopolyploid *S. anglica*, both parental repeats (from *S. maritima* and *S. alterniflora*) are present in most European populations investigated so far, which indicates that intergenomic locus homogenization does not seem to be occurring (Aïnouche et al. 2004; Baumel et al. 2001); however, the

homoeologous ratios may vary within and between populations, the "*maritima*-type" repeats being most frequently predominant (Huska et al. unpublished). Recent analyses from 454 Pyrosequencing in the parental hexaploid *Spartina* species reveal differential homogenization levels among the different parts of the 45S sub-repeats in this gene family (Boutte et al. unpublished). The *ITS* region and the coding (18S, 5.8S and 26S) regions of the transcription unit are relatively homogeneous, whereas substantial intragenomic heterogeneity in encountered in the External Transcribed Spacers (ETS) and Intergenic Spacers (IGS) reflecting the presence of various (paralogous and /or homoeologous) ribotypes within hexaploids.

A surprising finding from recent analyses in cotton is that nonindependent evolution of homoeologs is not restricted to multigenic families such as ribosomal DNA, but may also affect a large percentage of duplicated low-copy number genes. Using diagnostic homoeo-SNPs (Single Nucleotide Polymorphisms) from EST datasets, Salmon et al. (2010), and Flagel et al. (2012) found that up to 5 % of the polyploid transcriptome had experienced homoeologous gene conversion. More recently, (Page et al. 2013) used genome resequencing techniques conversion events that overlapped 113 genes.

One interesting dimension of these studies is the demonstration that nonreciprocal homoeologous exchanges have occurred in different lineages and at different rates, as opposed to only early during allopolyploid formation (Salmon et al. 2010), prior to reinforcement of strict bivalent pairing. About 50 % more homoeologous exchanges were detected in *G. hirsutum* than in *G. barbadense* (Flagel et al. 2012), indicating that rates of gene conversion may vary even among closely related species. These results lead to a number of questions regarding the functional and possibly adaptive consequences of such phenomena.

5.4 Evolution of Gene Expression Following Allopolyploidy

5.4.1 Evolution of Duplicate Gene Expression

Allopolyploidy has genome-wide effects on duplicate gene expression, ranging from equal and additive expression, relative to that of the diploid progenitors, to various forms of expression bias, silencing, and transgressive expression (Grover et al. 2012). In the absence of expression evolution (null hypothesis), allopolyploids are expected to exhibit an average of the expression levels of the corresponding parents, which usually is evaluated by comparison to a theoretical "mid-parent expression value" (Chen 2007), obtained *in silico* (from the expression level observed in the parental species) or experimentally (by analyzing an equimolar mix of parental RNA, which assumes a 50/50 parental contribution).

The advent of microarrays provided the first global perspective on transcriptome evolution, by comparing the total expression level of an allopolyploid to its parents and/or to the mid-parental expression value at several thousand loci. In cotton, Rapp et al. (2009) used a microarray platform designed against 40,430 unigenes to compare gene expression in two sets of parental diploids and their colchicine-doubled allopolyploid derivatives. Up to half of all genes were differentially expressed among diploids. Twelve patterns of differential expression for an allopolyploid and its diploid parents were distinguished, including situations of "parental expression dominance" (a phenomenon initially described in synthetic *Arabidopsis* allopolyploids by Wang et al. 2006) where the expression of the allopolyploid mimics that of one of the parents. Rapp et al. (2009) dissected this phenomenon and showed that when genes where expressed at mid-parent levels, this often was achieved via genome wide expression level dominance whereby gene expression was either up- or down- regulated to the level of one of the two parents.

Extending these analyses to the five natural allotetraploid *Gossypium* species (Flagel and Wendel 2010) found that about 34 % of genes investigated exhibited also this phenomenon. Interestingly, this study showed that expression level dominance became more balanced on an evolutionary timescale, with respect to the two parents. That is, the magnitude actually increased on an evolutionary timescale, but its directional bias disappeared. In contrast, transgressive up- and down-regulation was found to be more frequent among all extant allopolyploids than in the synthetic allotetraploid used in Rapp et al. (2009).

A similar trend was reported in *Spartina*. Here, transcriptome analyses were first performed using heterologous rice microarrays (Chelaifa et al. 2010a, b). Leaves from the hexaploid species *S. maritima* and *S. alterniflora* displayed 1,247 differentially expressed genes. Most were found to be up-regulated in *S. alterniflora.* The allopolyploid *S. anglica* displayed a moderate expression level dominance (similar to the expression levels observed in the maternal parent *S. alterniflora*) and nonadditive expression, including 118 transgressively expressed genes of which 101 were up-regulated in the allopolyploid. High throughput sequencing (Roche 454 and Illumina technologies) technologies have applied to the analysis of gene expression in *Spartina*, including transcriptome reference assemblies for the parental species ($\sim$17,000 annotated *Spartina* genes expressed in leaves and roots) (Ferreira de Carvalho et al. 2013b) and RNA-Seq analyses in the parents, F1 hybrid and allopolyploid (Boutte et al. unpublished). These analyses will soon provide a more complete picture of transcriptome evolution following allopolyploid speciation, including the possibility of distinguishing the contribution of each homeolog (see below).

5.4.2 Biased Expression of Homoeologs

Gossypium is one of the most thoroughly investigated polyploid system with regard to expression evolution of homoeologs. The first indication that polyploidy in *Gossypium* is accompanied by biased duplicate gene expression emerged a

decade ago in a study of 40 homoeologous gene pairs in different organs of *G. hirsutum* using SSCP-cDNA (Adams et al. 2003). Almost one-third of the genes revealed bias toward one homoeolog or the other, or only expression of one homoeolog, in at least one organ. Transcript levels for the two members of each gene pair varied by gene and, unexpectedly, by organ type. Especially noteworthy were genes that showed organ-specific, reciprocal silencing of alternative homoeologs; that is, one member of a duplicated gene pair displayed minimal to no transcription in some organs, whereas a reciprocal pattern was exhibited by the alternative homoeolog in other organ(s). In particular, floral organs showed dramatic expression patterns in this regard, with major differences among petals, stamens, and stigmas/styles (Adams et al. 2003, 2004).

Subsequent studies have confirmed these general conclusions and extended them to a genome wide scale. For example, Chaudhary et al. (2009) used high-resolution, genome-specific, mass-spectrometry technology to investigate relative expression levels of each homoeolog for 63 gene pairs in 24 tissues in naturally occurring allopolyploid cotton, a synthetic allopolyploid of the same genomic composition, and models of the diploid progenitor species. Results from a total of 2,177 successful expression assays indicated that 40 % of homeologs are transcriptionally biased in at least one stage of cotton development. Transcriptional subfunctionalization and 15 cases of probable neofunctionalization among eight tissues were encountered. Similarly, Udall et al. (2006) developed homoeolog-specific microarrays that utilized genome-diagnostic SNPs from ESTs generated from the two genomes of allopolyploid cotton and the diploids *G. arboreum* (A-genome) and *G. raimondii* (D-genome). Using leaf RNA, they found that 199 of 461gene pairs (43 %) deviated from equal expression. This microarray approach was refined and extended to a larger set of gene pairs ($n = 1{,}383$) by Flagel et al. (2008), who reported that 70 % of the genes in petals have biased homoeolog expression ratios, and that more of these genes are D-genome (39.5 %) than A-genome (30.5 %) biased. In addition, they found that the D-genome copies of 69 genes and the A-genome copies of 46 genes were silenced, collectively representing about 8 % of all genes. Remarkably, duplicate gene expression bias extends even to the level of the single-celled cotton fiber, as shown by Hovav et al. (2008a, b), who showed that 25–37 % of genes were significantly biased toward one of the two parental genomes at each developmental stage. Finally, and to place homoeolog expression evolution in a phylogenetic context, Flagel et al. (2012) explored duplicate gene expression in petals of five allopolyploid *Gossypium* species, demonstrating that all five species display an overall preference for D-genome expression (D-genome bias accounting for 54–60 % of genes with biased homoeolog expression), that the percentage of duplicate genes that are biased varies widely among species (from 48 to 88 %), and that the overall *magnitude* of bias (as opposed to simply whether or not a gene exhibits bias) similarly varies widely among species. This expression bias toward the D-genome was also encountered at the proteomic level (Hu et al. 2011). However, recently Rambani et al. (2014) studied duplicate gene expression in diploid and allopolyploid cottons using RNA seq technology, and found little directional preference in

gene expression bias, as well as a more modest level of expression level dominance than in earlier studies. Possible explanation for these partially conflicting results include differences in technologies and tissues employed, as well as environmental variation, as reported in coffee (Bardil et al. 2011).

5.4.3 Hybridization and Genome Duplication: Two Temporally Distinct Phases of Expression Evolution in Allopolyploids

The allopolyploid speciation process involves two steps that may be distinguished regarding their potential functional effects: Hybridization, which entails the reunion of two more or less divergent genomes and regulatory systems, and genome duplication with its attendant genomic redundancy and accompanying effects on gene expression and regulatory networks. It is useful to experimentally distinguish these two distinct aspects of allopolyploid speciation.

Spartina is one of few systems in which the F1 natural hybrid (*S. x townsendii*) that generated the neo-allopolyploid (*S. anglica*) remains extant4. Moreover, a second, independently formed hybrid (*S. x neyrautii*) from the same parental species represents a "natural replicate" of hybridization event, facilitating insight into the effects of combining two divergent genomes into a common nucleus. This framework has been useful for revealing the epigenetic and transcriptomic changes accompanying hybridization, including DNA methylation alterations triggered by hybridization in both *S. x townsendii* and *S. x neyrautii* (Salmon et al. 2005). Genome duplication does not entail significant additional methylation change, as *S. anglica* has inherited most of the changes observed in *S. x townsendii* but exhibits few new methylation alterations. As mentioned above, Parisod et al. (2009) have shown that an important fraction of these methylation changes affect regions flanking TEs, which agrees with the general view of methylation having evolved to control TEs in eukaryotic genomes (Slotkin and Martienssen 2007) and with the fact that no burst of transposition was detected following allopolyploid speciation in *Spartina* (Baumel et al. 2002a; Parisod et al. 2010). Transcriptome analyses using heterologous rice microarrays (Chelaifa et al. 2010b) revealed similar levels of nonadditive parental patterns of gene expression in both F1 hybrids *S. x townsendii* and *S. x neyrautii* (6.1 and 6.4 % of the analysed genes respectively), including parental (mostly maternal) gene expression dominance and transgressively expressed genes. However, the maternal expression dominance appeared more pronounced in *S. x townsendii* than in *S. x neyrautii.* About 8.7 % of the analyzed genes were found differentially expressed between these two F_1 hybrids and interestingly, most transgressively expressed genes were different, with genes upregulated in *S. x townsendii* being related to transport, detoxification, and stress, and genes upregulated in *S. x neyrautii* being related to cellular growth and development. The two independent hybridization events involving the same parental species then appear to have generated differential consequences in terms

of gene expression. The functions of these differentially expressed genes are consistent with the phenotypic differences between the two hybrids (see below). Genome duplication in *S. anglica* entailed additional transcriptome changes, consisting in the attenuation of the maternal dominance observed in the F_1 hybrid and an increased number of transgressively overexpressed genes (Chelaifa et al. 2010b). Thus, both hybridization and genome duplication appear to have important, though different effects on the *Spartina* transcriptome, occurring shortly after genome merger and polyploidization. For the first time, these decoupled effects were analysed during the allopolyploid speciation process, by comparing the actual (naturally formed) F_1 hybrid to its immediately derived allopolyploid that formed and survived in natural conditions.

In most allopolypoids, the actual F1 hybrid is generally not available, and the allopolyploid d origin via an F1 hybrid step may not be ascertained. Experimentally re-synthesized hybrids using the present-day parental representatives may be an alternative way to dissect the respective effects of hybridization and genome duplication and to infer the changes occurred since the polyploidization event by comparison to the natural allopolyploid, an opportunity that is offered by the 1–2 my old allotetraploid *Gossypium* species which evolved under different natural and artificial (domestication) conditions. Interestingly, the findings in *Spartina* mentioned above seem to parallel the conclusions emerging from similar comparisons involving natural, more or less recent allopolyploids and/or synthetic F_1 hybrids in *Gossypium* (Flagel et al. 2008; Flagel and Wendel 2010), *Senecio* (Hegarty et al. 2006) and *Tragopogon* (Buggs et al. 2011).

In *Gossypium*, Flagel et al. (2008) used a microarray platform capable of measuring homoeolog-specific expression, to compare petal expression changes between the diploid parental species (A-genome representative *G. arboreum* and D-genome representative *G. raimondii*), a synthetic F1 hybrid and the natural (1–2 my old) allotetraploid *G. hirsutum*. For petal tissues, expression bias was found favoring the parental D-genome in the F1 hybrid, which became further enhanced in the natural allotetraploid. These data showed that a significant fraction of expression bias found in allotetraploids likely is initiated by genome merger per se. Long-term evolutionary processes build on this initial genome-wide expression modification, thus implicating two temporally distinct phases of expression evolution following allopolyploidization. (Flagel and Wendel 2010) extended the scope of these findings to a diverse collection of natural allopolyploid species, further refining our temporal perspective on expression evolution and revealing extraordinary variation in the rate of expression evolution during radiation of an allopolyploid lineage. They also show aspects of expression evolution that are shared among the five natural allotetraploid cotton species, as well as those that are distinct from expression changes in recently formed synthetic inter-genomic hybrids. Specifically, the over-representation of the D-genome bias is largely reversed among all five natural allopolyploids, both at the homoeolog level and among total gene expression profiles. That is, over evolutionary time, the allotetraploids begin to assume roughly equivalent numbers of A- and D-dominant states. Interestingly, it is not the *magnitude* of genomic dominance that is altered

by time, but its *direction*. Thus, it would appear that the allopolyploid genomes have adjusted, during 1–2 million years of evolution, to more equally utilize the transcriptomes of the two co-resident genomes. Chaudhary et al. (2009) also found, in various tissues, that genome merger has the largest impact on biased expression of homoeologs along the pathway to polyploidy in cotton, and that the majority of these alterations are caused by *cis*-regulatory divergence between the diploid progenitors. Finally, when comparing the interspecific F1 hybrid, synthetic and natural allopolyploid cottons using RNA-Seq analyses from leaf transcriptomes, Yoo et al. (2013) also observed biased transcriptome preference for the D genome in the synthetic allopolyploid, whereas the direction was reversed in the hybrid and natural allopolyploids. Radical alterations in homoeolog bias and transcriptome mimicry accompany the initial merger of two diverged diploid genomes, suggesting a combination of *cis*- and *trans*-regulatory and epigenetic interactions and modifications that propagate throughout the transcriptome network. The magnitude of homoeolog bias and expression level dominance temporally increases from hybridization through polyploid evolution. These observations collectively suggest the general conclusions that initial conditions matter, but that natural selection ultimately reconciles the regulatory mismatches caused by genomic merger, while new gene expression space is generated.

5.5 Phenotypic Novelty and Adaptive Consequences of Allopolyploidy

The question naturally arises as to whether the kinds of genomic and transcriptomic changes reviewed here have stimulated novel phenotypes (ecological, physiological, morphological) and adaptation. A voluminous literature in plants documents the frequency of polyploids in various habitats, their morphological and physiological attributes, and their ecological success relative to diploids (Grant 1971; Soltis and Soltis 2000; Stebbins 1947, 1950). One generalization that has emerged is that polyploidy often is associated with broader ecological amplitude and novel evolutionary opportunity, often suggested to be mediated by the increased "buffering" capacity afforded by duplicated genes and the enhanced vigor resulting from the "fixed heterozygosity" of their duplicated genomes. We might now rephrase this explanation to encompass a network perspective, one that recognizes the vastly increased combinatorial possibilities for regulation and evolution enabled by a suddenly duplicated complement of genes and merged regulatory systems.

5.5.1 Ecological Novelty

In both *Spartina* and *Gossypium*, allopolyploidy led to the apparent invasion of a new ecological niche. In considering the Pleistocene origin of allopolyploid cotton, Fryxell (1965, 1979) noted that in contrast to the majority of diploid species,

allopolyploid species typically occur in coastal habitats, at least those forms that arguably are truly wild. Thus, among the allopolyploid species, two are completely restricted to near coastlines, in that they are island endemics (*G. darwinii* and *G. tomentosum*), and for two others (*G. barbadense* and *G. hirsutum*), wild forms occur disparately in littoral habitats ringing the Gulf of Mexico, northwest South America, and distant Pacific Islands. The capacity for oceanic dispersal in *Gossypium* (Fryxell 1965, 1979; Stephens 1958, 1966) was associated at the allopolyploid level with specialization for establishment in coastal communities. Fryxell (1965, 1979) forwarded the tantalizing suggestion that following initial formation, adaptation of the newly evolved allopolyploid to littoral habitats enabled it to exploit the fluctuating sea levels that characterized the Pleistocene. This ecological innovation is envisioned to have not only permitted the initial establishment of the nascent polyploid lineage, but is also suggested to have provided a means for the rapid dispersal of the salt-water tolerant seeds. By this means, perhaps, the mobile shorelines of the Pleistocene facilitated exploitation of a new ecological niche, and hence colonization of the New World tropics. It is tempting to correlate the functional flexibility resulting from the immediate gene expression changes presented in the above sections, to these long-term consequences on species ecology and distribution. In this respect, it is of interest to consider the results of Dong and Adams (2011), who reported that abiotic stress conditions had large effects on duplicate gene expression, with the effects varying by gene, stress and organ type. Perhaps wholesale gene duplication gave rise to the possibility of the exploration of new ecological niches characterized by added salinity, fluctuating water levels, or some other feature of coastal habitats.

Ecological implications of allopolyploidy are particularly prominent in *Spartina,* as some species play an important ecological role in the sedimentary dynamics of salt marshes, where the plants are considered to be “ecosystem engineers” (Crooks 2002). The ecological range of the neo-allododecaploid *S. anglica* along the shore is larger than either of its parents. *S. anglica* tolerates several hours of immersion at high tides, and thus is able to occupy a vacant niche as a pioneer species in the low tide zone. This species may accrete large volumes of tidal sediments, making the habitat more terrestrial, and allowing colonization by other salt marsh plant species, which modifies the physical structure of intertidal coastal zones.

Physiological and anatomical adaptations are important components of *Spartina* ecology and distribution (Maricle et al. 2006, 2009). As observed in many polyploids (Otto 2007), stomatal cell size increases with ploidy level in *Spartina* (Kim et al. 2010; Marchant 1967), which may affect photosynthetic rates (Warner and Edwards 1993). The larger ecological amplitude of the allopolyploid *S. anglica* is associated with increased tolerance to highly reducing and sulfidic sediment conditions. This increased tolerance may explain the ability of *S. anglica* to colonize low-marsh zones (Maricle et al. 2006). Survival of *S. anglica* in anoxic sediments likely is facilitated by its particular ability to develop aerenchyma systems that supply the submerged plants with atmospheric oxygen and efficiently transport oxygen to the roots (Maricle and Lee 2002). *Spartina anglica* displays

enhanced mechanisms to transport O_2 and exhibits five-time higher H_2S removal than its progenitor species *S. alterniflora* (Lee 2003). Interestingly, genes that appeared transgressively upregulated following hybridization (in *S. x townsendii*) and genome duplication (in *S. anglica*) were related to transport, detoxification and stress tolerance (Chelaifa et al. 2010b).

5.5.2 Phenotypic Novelty

Hybridization between *S. alterniflora* and *S. maritima* had very different morphological consequences in the two independent events that occurred in England (S. *x townsendii*) and France (*S. x neyrautii*), even though these hybridization events involved crosses in the same direction (*S. alterniflora* being the maternal genome donor). *Spartina* x *neyrautii* has shorter spikelets and is distinctly more slender than *S. x ownsendii*, which has longer fleshy leaves, resembling more closely the maternal parent *S. alterniflora,* whereas *S. x townsendii* has intermediate morphological features between *S. maritima* and *S. alterniflora* (Mobberley 1956).The phenotypic differences between these two F_1 hybrids of similar genetic origin are puzzling, and may reflect differential effects of "genomic shock." *Spartina x townsendii* is almost indistinguishable from its allopolyploid derivative *S. anglica*; moreover, the latter species exhibits large phenotypic plasticity (Thompson et al. 1991).

In *Gossypium*, allopolyploidy provided raw material for both natural and artificial selection, and this had critical impact on the development of agronomically advanced cultivars of cotton. Although four separate species of *Gossypium* were independently domesticated for their seed hairs, the characteristic that attracted the attention of the earliest domesticators, the seed "lint" itself, evolved only once in the progenitor the A-genome diploids (Wendel et al. 2012). Applequist et al. (2001) generated growth curves for trichomes from cultivated and wild diploid and allopolyploid species, and demonstrated that the evolution of an extended primary wall elongation occurred in the ancestor of wild A-genome cotton prior to domestication and in Africa. Follow-up comparative expression profiling experiments (Hovav et al. 2008a, b; Rapp et al. 2010) further identify some of the metabolic pathways that were modified to enable this evolution in fiber properties. These results lead to the fascinating implication that domestication of the New World allopolyploid cottons (which contain an A-genome, in addition to a D-genome) that presently dominate cotton agriculture worldwide was first precipitated by developmental and physiological transformations that occurred hundreds of thousands of years ago in a different hemisphere.

Because fibers from all D-genome diploids are short and nonspinable, it is particularly interesting that fiber from the cultivated (New World) allopolyploids is agronomically superior to that of the cultivated A-genome diploids; in this sense, the cultivated allotetraploid fiber morphology is "nonadditive," or perhaps "heterotic." A number of studies have noted this point (Jiang et al. 1998; Paterson

2005; Wright et al. 1998), suggesting that allopolyploidization provided novel opportunities for agronomic improvement. A recent meta-analysis (Rong et al. 2007) of a large number of QTL studies in allopolyploid cotton leads to a general picture consistent with this interpretation; more loci affecting fiber yield and quality traits are found in the D_T ($n = 221$) than the A_T ($n = 184$) genome, possibly explaining the superiority of the lint of the allopolyploids relative to the A-genome diploids. Support for this speculation that "recruitment" of D-genome genes has been important in enabling the development of advanced allotetraploid cultivars is also emerging from comparative expression profiling studies (Hovav et al. 2008a, b; Rapp et al. 2010), which reveal in exquisite detail the thousands of gene expression differences that distinguish wild from domesticated cotton fiber development. Finally, similar implications emanate from genetic mapping experiments, where it has been shown that for 535 genes implicated in cotton fiber development, more transcription factors were from D_T than the A_T genome, whereas the reverse was true for fiber development genes (Xu et al. 2010). These data are interpreted to suggest that the D-genome ancestor provided some of the key transcription factors that regulate the expression of fiber genes donated by the ancestral A-genome parent. Taken together, these studies may provide actual genetic evidence for a speculation forwarded 75 years ago by Harland (1936), who stated *If as a consequence of polyploidy a large number of genes become duplicated, and the characters governed by such genes are of importance to the species, one of the members may mutate, leaving the character unimpaired, with the further possibility that the mutation may be of benefit to the species*. An exciting prospect is that in the near future we will develop a deeper understanding of the nature of these genes, the molecular genetic meaning of Harland's invocation of the word "mutation," and their effects on the developmental networks that underlie altered morphology and agronomic improvement.

In conclusion, the well-established framework now available for recent allopolyploids such as *Spartina* and *Gossypium,* have offered unique opportunities to explore the phenotypic, ecological and genomic consequences of genome merger and duplication. The increased knowledge that has accumulated on various polyploid systems has revealed a large range of possible responses to allopolyploidy. *Spartina* and *Gossypium* have in common several features in this regard: In both systems, allopolyploidy was not accompanied by rapid restructuring of the parental genomes, nor bursts of transposition. Gene expression alteration in both allopolyploids is massively altered, but separately by the three temporally distinct phases of genome merger: hybridization, whole genome doubling, and subsequent evolution. The various mechanisms underlying gene expression evolution in allopolyploids are far from being fully understood, and much remains to be learned regarding the effects of these changes on the transcriptional, translational, and metabolic networks that ultimately lead to adaptation and novel phenotypes. The increasing application of ever-more sophisticated 'omics' technologies to these and other natural systems promises rapid progress regarding these central questions in the upcoming years.

Acknowledgments The Ainouche and Wendel labs wish to acknowledge the support of funding agencies for much of the research summarized here. These include the Partner University Fund, CNRS, University of Rennes 1 (supporting the International Associated Laboratory "ECOGEN: Ecological Genomics of Polyploidy"), Conseil Regional de Bretagne (France), Cotton Incorporated, and the National Science Foundation Plant Genome Program (USA). We also wish to thank the many students, post-doctoral research associates, and colleagues, who have contributed to the many studies described herein over the past 25 years.

References

Abbott RJ, Lowe AJ (2004) Origins, establishment and evolution of new polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. Biol J Linn Soc 82:467–474

Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Nat Acade Sci USA 100:4649–4654

Adams KL, Percifield R, Wendel JF (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics 168:2217–2226

Adams KL, Wendel JF (2005) Novel patterns of gene expression in polyploid plants. Trends Genet 21:539–543

Aïnouche ML, Baumel A, Salmon A (2004) *Spartina anglica* C.E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. Biol J Linn Soc 82:475–484

Ainouche ML, Chelaifa H, Ferreira de Carvalho J, Bellot S, Ainouche AK, Salmon A (2012) Polyploid evolution in *Spartina*: dealing with highly redundant hybrid genomes. In: Soltis PS, Soltis DE (eds) Polyploidy and genome evolution. Springer, Berlin, pp 225–243

Ainouche ML, Fortune PM, Salmon A, Parisod C, Grandbastien MA, Fukunaga K, Ricou M, Misset MT (2009) Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). Biol Invasions 11:1159–1173

Ainouche ML, Jenczewski E (2010) Focus on polyploidy. New Phytol 186:1–4

Applequist WL, Cronn RC, Wendel JF (2001) Comparative development of fiber in wild and cultivated cotton. Evol Dev 3:3–17

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Arnaud D, Chelaifa H, Jahier J, Chalhoub B (2013) Reprogramming of gene expression in the genetically stable bread allohexaploid wheat. In: Chen ZJ, Birchler JA (eds) Polyploid and hybrid genomics. Wiley, Oxford, pp 195–211

Bardil A, de Almeida JD, Combes MC, Lashermes P, Bertrand B (2011) Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. New Phytol 192:760–774

Baumel A, Ainouche M, Kalendar R, Schulman AH (2002a) Retrotransposons and genomic stability in populations of the young allopolyploid species *Spartina anglica* C.E. Hubbard (Poaceae). Mol Biol Evol 19:1218–1227

Baumel A, Ainouche M, Misset M, Gourret J, Bayer R (2003) Genetic evidence for hybridization between the native *Spartina maritima* and the introduced *Spartina alterniflora* (Poaceae) in South-West France: Spartina x neyrautii re-examined. Plant Syst Evol 237:87–97

Baumel A, Ainouche ML, Bayer RJ, Ainouche AK, Misset MT (2002b) Molecular phylogeny of hybridizing species from the genus *Spartina* Schreb. (Poaceae). Mol Phylogenet Evol 22:303–314

Baumel A, Ainouche ML, Levasseur JE (2001) Molecular investigations in populations of *Spartina anglica* C.E. Hubbard (Poaceae) invading coastal Brittany (France). Mol Ecol 10:1689–1701

Birchler JA, Veitia RA (2010) The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. New Phytol 186:54–62
Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res 13:137–144
Brubaker CL, Bourland FM, Wendel JF (1999a) The origin and domestication of cotton. In: Smith CW, Cothren JT (eds) Cotton; origin, history, technology and production. Wiley, New York, pp 3–31
Brubaker CL, Paterson AH, Wendel JF (1999b) Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42:184–203
Brubaker CL, Wendel JF (1994) Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*: Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). Am J Bot 81:1309–1326
Buggs RJ, Zhang L, Miles N, Tate JA, Gao L, Wei W, Schnable PS, Barbazuk WB, Soltis PS, Soltis DE (2011) Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol 21:551–556
Chaudhary B, Flagel L, Stupar RM, Udall JA, Verma N, Springer NM, Wendel JF (2009) Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). Genetics 182:503–517
Chelaifa H, Mahe F, Ainouche M (2010a) Transcriptome divergence between the hexaploid salt-marsh sister species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Mol Ecol 19:2050–2063
Chelaifa H, Monnier A, Ainouche M (2010b) Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina x townsendii* and *Spartina anglica* (Poaceae). New Phytol 186:161–174
Chen J (2007) Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol 58:377–406
Chen ZJ, Birchler JA (2013) Polyploid and hybrid genomics. Wiley, New York
Chester M, Gallagher JP, Symonds V, Veruska Cruz da Silva A, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE (2012) Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). Proc Nat Acad Sci USA 109:1045–1057
Comai L (2005) The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–846
Cronn R, Small RL, Wendel JF (1999) Duplicated genes evolve independently following polyploid formation in cotton. Proc Nat Acad Sci USA 96:14406–14411
Cronn RC, Small RL, Haselkorn T, Wendel JF (2002) Rapid diversification of the cotton genus (*Gossypium* : Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot 89:707–725
Crooks J (2002) Characterizing ecosystem-level consequences of biological invasions: the role of ecosystem engineers. Oikos 97:153–166
Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–749
De Smet R, Van de Peer Y (2012) Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr Opin Plant Biol 15:168–176
Dobzhansky T (1937) Genetics and the origins of species. Columbia University Press, New York
Dong S, Adams KL (2011) Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. New Phytol 190:1045–1057
Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42:443–461
Elder JF, Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. Quart Rev Biol 70:297–320
Endrizzi JE, Turcotte EL, Kohel RJ (1985) Genetics, cytology and evolution of *Gossypium*. Adv Genet 23:271–375

Ferreira de Carvalho J, Chelaifa H, Boutte J, Poulain J, Couloux A, Wincker P, Bellec A, Fourment J, Berges H, Salmon A, Ainouche M (2013a) Exploring the genome of the salt-marsh *Spartina maritima* (Poaceae, Chloridoideae) through BAC end sequence analysis. Plant Mol Biol 83:591–606

Ferreira de Carvalho J, Poulain J, Da Silva C, Wincker P, Michon-Coudouel S, Dheilly A, Naquin D, Boutte J, Salmon A, Ainouche M (2013b) Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Heredity (Edinb) 110:181–193

Ferris C, King RA, Gray AJ (1997) Molecular evidence for the maternal parentage in the hybrid origin of *Spartina anglica* C.E. Hubbard. Mol Ecol 6:185–187

Flagel L, Udall J, Nettleton D, Wendel J (2008) Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. BMC Biol 6:11

Flagel L, Wendel JF, Udall J (2012) Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genom 13:302. doi:10.1186/1471-2164-13-302

Flagel LE, Wendel JF (2010) Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol 186:184–193

Flagel LE, Wendel JF, Udall JA (2011) Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics in review

Fortune PM, Schierenbeck K, Ayres D, Bortolus A, Catrice O, Brown S, Ainouche ML (2008) The enigmatic invasive *Spartina densiflora*: a history of hybridizations in a polyploidy context. Mol Ecol 17:4304–4316

Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML (2007) Evolutionary dynamics of *Waxy* and the origin of hexaploid *Spartina* species (Poaceae). Mol Phylogenet Evol 43:1040–1055

Foucaud (1897) Un *Spartina* inédit. Ann Soc Sci Nat Char Inf 32:220–222

Fryxell PA (1965) Stages in the evolution of *Gossypium*. Adv Front Plant Sci 10:31–56

Fryxell PA (1979) The natural history of the cotton tribe. Texas A&M University Press, College Station

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007) Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. Plant Cell 19:3403–3417

Grant V (1971) Plant speciation. Columbia University Press, New York

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2004) Incongruent patterns of local and global genome size evolution in cotton. Genome Res 14:1474–1482

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2007) Microcolinearity and genome evolution in the *AdhA* region of diploid and polyploid cotton (*Gossypium*). Plant J 50:995–1006

Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. Am J Bot 99:312–319

Hanson RE, Islam-Faridi MN, Crane CF, Zwick MS, Czeschin DG, Wendel JF, Mcknight TD, Price HJ, Stelly DM (1999) Ty1-*copia*-retrotransposon behavior in a polyploid cotton. Chromosome Res 8:73–76

Hanson RE, Zhao X-P, Islam-Faridi MN, Paterson AH, Zwick MS, Crane CF, McKnight TD, Stelly DM, Price HJ (1998) Evolution of interspersed repetitive elements in *Gossypium* (Malvaceae). Am J Bot 85:1364–1368

Harland SC (1936) The genetical conception of the species. Camb Philos Society Biol Rev 11:83–112

Hawkins JS, Hu G, Rapp RA, Grafenberg JL, Wendel JF (2008) Phylogenetic determination of the pace of transposable element proliferation in plants: copia and LINE-like elements in *Gossypium*. Genome 51:11–18

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261

Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009) Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc Nat Acad Sci USA 106:17811–17816

Hegarty MJ, Abbott RJ, Hiscock SJ (2012) Allopolyploid speciation in action: the origins and evolution of *Senecio cambrensis*. In: Soltis PS, Soltis DE (eds) Polyploidy and genome evolution. Springer, Berlin, pp 245–270

Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ (2006) Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. Curr Biol 16:1652–1659

Hendrix B, Stewart JM (2005) Estimation of the nuclear DNA content of *Gossypium* species. Ann Bot 95:789–797

Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF (2008a) Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. Genetics 179:1725–1733

Hovav R, Udall J, Chaudhary B, Flagel L, Rapp R, Wendel J (2008b) Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proc Nat Acad Sci USA 105:6191

Hu G, Hawkins JS, Grover CE, Wendel JF (2010) The history and disposition of transposable elements in polyploid *Gossypium*. Genome 53:599–607

Hu G, Houston NL, Pathak D, Schmidt L, Thelan JJ, Wendel JF (2011) Genomically biased accumulation of seed storage proteins in allopolyploid cotton. Genetics 189:1103−1115

Hubbard JCE (1968) Grasses, 2nd edn. Penguin Books, London

Huskin C (1930) The origin of *S. x townsendii*. Genetica 12:531–538

Hutchinson JB (1951) Intra-specific differentiation in *Gossypium hirsutum*. Heredity 5:161–193

Hutchinson JB (1954) New evidence on the origin of the old world cottons. Heredity 8:225–241

Hutchinson JB, Silow RA, Stephens SG (1947) The evolution of *Gossypium* and the differentiation of the cultivated cottons. Oxford University Press, London

Jackson S, Chen ZJ (2010) Genomic and expression plasticity of polyploidy. Curr Opin Plant Biol 13:153–159

Jiang C, Wright R, El-Zik K, Paterson A (1998) Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proc Nat Acad Sci USA 95:4419–4424

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, de Pamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100

Jovet P (1941) Notes systématiques et écologiques sur les Spartines du Sud-Ouest. Bulletin de la Société Botanique de France 88:115–123

Kim SM, Rayburn AL, Lee DK (2010) Genome size and chromosome analysis in prairie cordgrass (*Spartina pectinata* L.). Crop Sci 50:2277–2282

Kovarik A, Matyasek R, Lim K, Skalická K, Koukalová B, Knapp S, Chase M, Leitch A (2004) Concerted evolution of 18–5.8–26S rDNA repeats in *Nicotiana* allotetraploids. Biol J Linn Soc 82:615–625

Lee R (2003) Physiological adaptations of the invasive cordgrass *Spartina anglica* to reducing sediments: rhizome metabolic gas fluxes and enhanced O_2 and H_2S transport. Mar Biol 143:9–15

Leitch IJ, Bennett MD (2004) Genome downsizing in polyploid plants. Biol J Linn Soc 82:51–663

Lewis WH (1980) Polyploidy: biological relevance. Plenum Press, New York

Lim KY, Soltis DE, Soltis PS, Tate J, Matyasek R, Srubarova H, Kovarik A, Pires JC, Xiong Z, Leitch AR (2008) Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae). PLoS ONE 3:e3353

Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF (2001a) Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome 44:321–330

Liu Q, Brubaker CL, Green AG, Marshall DR, Sharp PJ, Singh SP (2001b) Evolution of the *FAD*2-1 fatty acid desaturase 5′ UTR intron and the molecular systematics of *Gossypium* (Malvaceae). Am J Bot 88:92–102

Maere S, De Bodt S, Raes J, Castneuf J, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Nat Acad Sci USA 102:5454–5459

Marchant C (1963) Corrected chromosome numbers for *Spartina x townsendii* and its parent species. Nature 199:929

Marchant C (1967) Evolution in *Spartina* (Gramineae): I. The history and morphology of the genus in Britain. Bot J Linn Soc 60:1–24

Marchant C (1968) Evolution in *Spartina* (Graminae). III Species chromosome numbers and their taxonomic signifiance. Bot J Linn Soc 60:411–417

Maricle B, Crosier J, Bussiere B, Lee R (2006) Respiratory enzyme activities correlate with anoxia tolerance in saltmarsh grasses. J Exp Mar Biol Ecol 337:30–37

Maricle B, Koteyeva N, Voznesenskaya E, Thomasson J, Edwards G (2009) Diversity in leaf anatomy, and stomatal distribution and conductance, between salt marsh and freshwater species in the C4 genus *Spartina* (Poaceae). New Phytol 184:216–233

Maricle B, Lee R (2002) Aerenchyma development and oxygen transport in the estuarine cordgrasses *Spartina alterniflora* and *S. anglica*. Aquat Bot 74:109–120

Mayr E (1942) Systematics and the origin of species. Columbia University Press, New York

McClintock B (1984) The significance of responses of the genome to challenge. Science 226:792–801

Mobberley D (1956) Taxonomy and distribution of the genus *Spartina*. Iowa State Coll J Sci 30:471–574

Ohno S (1970) Evolution by gene duplication. Springer, New York

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA (2003) Understanding mechanisms of novel gene expression in polyploids. Trends Genet 19:141–147

Otto SP (2007) The evolutionary consequences of polyploidy. Cell 131:452–462

Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet 34:401–437

Ozkan H, Levy AA, Feldman M (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. Plant Cell 13:1735–1747

Page JT, Huynh MD, Lichty ZS, Grupp K, Stelly D, Hulse A, Ashrafi H, van Deynze A, Wendel JF, Udall JA (2013) Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. G3: genes. Genomes Genet 3:1809–1818

Pandit M, Pocock M, Kunin W (2011) Ploidy influences rarity and invasiveness in plants. J Ecol 99:1108–1115

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186:37–45

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien MA, Ainouche M (2009) Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. New Phytol 184:1003–1015

Paterson AH (2005) Polyploidy, evolutionary opportunity, and crop adaptation. Genetica 123:191

Paterson AH, Wang XY, Li JP, Tang HB (2012) Ancient and recent polyploidy in Monocots. In: Soltis PS, Soltis DE (eds) Polyploidy and genome evolution. Springer, Berlin, pp 93–108

Percy RG, Wendel JF (1990) Allozyme evidence for the origin and diversification of *Gossypium barbadense* L. Theor Appl Genet 79:529–542

Rambani A, Page JT, Udall JA (2014) Polyploidy and the petal transcriptome of *Gossypium*. BMC Plant Biol 14:3

Ramsey J, Schemske DW (1998) Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu Rev Ecol Syst 29:467–501

Rapp R, Haigler C, Flagel L, Hovav R, Udall J, Wendel J (2010) Gene expression in developing fibres of Upland cotton (*Gossypium hirsutum* L.) was massively altered by domestication. BMC Biol 8:139

Rapp RA, Udall JA, Wendel JF (2009) Genomic expression dominance in allopolyploids. BMC Biol 7:18

Renny-Byfield S, Ainouche M, Leitch IJ, Lim KY, Le Comber SC, Leitch AR (2010) Flow cytometry and GISH reveal mixed ploidy populations and *Spartina* nonaploids with genomes of *S. alterniflora* and *S. maritima* origin. Ann Bot 105:527–533

Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park C, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Williams-Coplin TD, Wing RA, Wright RJ, Zhao X, Zhu L, Paterson AH (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166:389–417

Rong J, Feltus EA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, Smith CW, Gannaway JR, Wendel JR, Paterson AH (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176:2577–2588

Rong J, Feltus FA, Liu L, Lin L, Paterson AH (2010) Gene copy number evolution during tetraploid cotton radiation. Heredity (Edinb) 105:463–472

Salmon A, Ainouche ML, Wendel JF (2005) Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). Mol Ecol 14:1163–1175

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF (2010) Homoeologous nonreciprocal recombination in polyploid cotton. New Phytol 186:123–134

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong JK, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol Biol Evol 20:633–643

Skalicka K, Lim KY, Matyasek R, Matzke M, Leitch AR, Kovarik A (2005) Preferential elimination of repeated DNA sequences from the paternal, *Nicotiana tomentosiformis* genome donor of a synthetic, allotetraploid tobacco. New Phytol 166:291–303

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285

Small RL, Wendel JF (1999) The mitochondrial genome of allotetraploid cotton (*Gossypium* L.). J Hered 90:251–253

Small RL, Wendel JF (2002) Differential evolutionary dynamics of duplicated paralogous Adh loci in allotetraploid cotton (*Gossypium*). Mol Biol Evol 19:597–607

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol Evol 14:348–352

Soltis DE, Soltis PS, Pires JC, Kovarik A, Tate JA, Mavrodiev E (2004) Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. Biol J Linn Soc 82:485–501

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Nat Acad Sci USA 97:7051–7057

Soltis PS, Soltis DE (2012) Polyploidy and genome evolution. Springer, Berlin

Stebbins GL (1947) Types of polyploids: their classification and significance. Adv Genet 1:403–429

Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York

Stephens SG (1958) Salt water tolerance of seeds of *Gossypium* species as a possible factor in seed dispersal. Amer Nat 92:83–92

Stephens SG (1966) The potential for long range oceanic dispersal of cotton seeds. Amer Nat 100:199–210

Strong DR, Ayres DR (2013) Ecological and evolutionary misadventures of *Spartina*. Ecol Evol Syst 44:389–410

Szadkowski E, Eber F, Huteau V, Lodé M, Huneau C, Belcram H, Coriton O, Manzanares-Dauleux MJ, Delourme R, King GJ, Chalhoub B, Jenczewski E, Chèvre AM (2010) The first meiosis of resynthesized *Brassuca napus*, a genome blender. New Phytol 186:102–112

Tate JA, Joshi P, Soltis KA, Soltis PS, Soltis DE (2009) On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). BMC Plant Biol 9:80

te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubesova M, Pysek P (2012) The more the better? The role of polyploidy in facilitating plant invasions. Ann Bot 109:19–45

Thompson JD, McNeilly T, Gray AJ (1991) Population variation in *Spartina anglica* C.E. Hubbard. I. Evidence from a common garden experiment. New Phytol 117:115–128

Treier UA, Broennimann O, Normand S et al (2009) Shift in cytotype frequency and niche space in the invasive plant *Centaurea maculosa*. Ecology 90:1366–1377

Udall JA, Swanson JM, Nettleton D, Percifield RJ, Wendel JF (2006) A novel approach for characterizing expression levels of genes duplicated by polyploidy. Genetics 173:1823–1827

Wang JL, Tian L, Lee HS, Wei NE, Jiang HM, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, Chen ZJ (2006) Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. Genetics 172:507–517

Warner DA, Edwards GE (1993) Effects of polyploidy on photosynthesis. Photosynth Res 35:135–147

Wendel JF (1989) New world tetraploid cottons contain old world cytoplasm. Proc Nat Acad Sci USA 86:4132–4136

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol 42:225–249

Wendel JF, Brubaker CL, Alvarez JP, Cronn RC, Stewart JM (2009) Evolution and natural history of the cotton genus. In: Paterson AH (ed) Genomics of cotton, plant genetics and genomics, crops and models 3. Springer, New York, pp 3–22

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

Wendel JF, Doyle JJ (2005) Polyploidy and evolution in plants. In: Henry RJ (ed) Plant diversity and evolution. CABI Publishing, Wallington, pp 97–117

Wendel JF, Flagel LE, Adams KL (2012) Jeans, genes, and genomes: cotton as a model for studying polyploidy. In: Soltis PS, Soltis DE (eds) Polyploidy and genome evolution. Springer, Berlin, pp 181–207

Wendel JF, Schnabel A, Seelanan T (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc Nat Acad Sci USA 92:280–284

Wright RJ, Thaxton PM, El-Zik KM, Paterson AH (1998) D-Subgenome bias of *Xcm* resistance genes in tetraploid *Gossypium* (cotton) suggests that polyploid formation has created novel avenues for evolution. Genetics 149:1987–1996

Xu Z, Yu JZ, Cho J, Yu J, Kohel RJ, Percy RG (2010) Polyploidization altered gene functions in cotton (*Gossypium* spp.). PLoS ONE 5:e14351

Yannic G, Baumel A, Ainouche M (2004) Uniformity of the nuclear and chloroplast genomes of *Spartina maritima* (Poaceae), a salt-marsh species in decline along the Western European Coast. Heredity (Edinb) 93:182–188

Yoo MJ, Szadkowski E, Wendel JF (2013) Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) 110:171–180

Zozomova-Lihova J, Krak K, Mandakova T, Shimizu KK, Spaniel S, Vit P, Lysak MA (2014) Multiple hybridization events in *Cardamine* (Brassicaceae) during the last 150 years: revisiting a textbook example of neoallopolyploidy. Ann Bot doi:10.1093/aob/mcu012

Chapter 6
Evolutionary Divergence in Human Versus Mouse Innate Immune Gene Regulation and Function

Ronan Kapetanovic, Juliana K. Ariffin and Matthew J. Sweet

Abstract Gene and/or pathway conservation across species implies essential functions for that gene or pathway in a particular biological response. Such conservation is particularly important in studies of model organisms, where one wishes to infer biology in one species, based on studies in another. In this respect, murine studies have been particularly informative for generating insights into human physiological and pathophysiological processes. Some biological systems are particularly susceptible to evolutionary change, however, as is the case with the innate immune system that must co-evolve with rapidly evolving pathogens. An understanding of evolutionary conservation and divergence in innate immune pathways across species can provide insights into both species-specific immune responses that are likely to be important for host defence, as well as limitations of model organisms for studies of innate immune processes relevant to human disease. In this chapter, we discuss genetic differences in human versus mouse innate immunity, as well as specific mechanisms that contribute to such differences. While we provide a broad overview of several innate immune gene families, we focus in more detail on the Toll-like receptor 4 pathway, which is involved in sensing lipopolysaccharide from Gram-negative bacteria.

6.1 Introduction

6.1.1 An Overview of Innate Immunity

The innate immune system coordinates a rapid inflammatory response upon sensing danger. This has been most widely studied in the context of infections, where activation of innate immunity enables elimination of the invading

R. Kapetanovic (✉) · J. K. Ariffin · M. J. Sweet
Division of Molecular and Cell Biology, Institute for Molecular Bioscience, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia
e-mail: r.kapetanovic@imb.uq.edu.au

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_6,

microorganism(s) or, if overwhelmed, the containment of the infection until an antigen-specific adaptive immune response can be mounted. Innate immunity research first came into prominence in the late 1800s, with the description of phagocytosis, a process by which innate immune cells take up and degrade particulate matter. Nobel Prize recipient Ilya Mechnikov detailed this process and was the first to coin the term phagocytosis (Gordon 2008), but the existence of this cellular pathway was probably first identified by William Osler several years earlier (Ambrose 2006). After this spotlight on innate immunity research, the adaptive immune system became the focus of much of the immunological research community throughout the twentieth century. The apparent lack of interest in innate immunity perhaps stemmed, at least in part, from a lack of knowledge about the molecular processes leading to innate immune activation. Innate immunity was thus widely regarded as a 'non-specific' form of host defence. This view was finally challenged by Charles Janeway in 1989, who proposed that the innate immune system must possess specific pattern recognition receptors (PRRs) to recognise pathogen-associated molecular patterns (PAMPs) present on invading microorganisms (Janeway 1989). Janeway's hypothesis was confirmed with the discovery by Jules Hoffmann's team of the essential role of the toll receptor in host defence to microbial challenge in drosophila (Lemaitre et al. 1996). This work, for which he received the Nobel Prize in 2011, along with Bruce Beutler and Ralph Steinman, ultimately led to the discovery of several families of PRRs, such as the Toll-like receptors (TLRs) and Nod-like receptors (NLRs) that enable the innate immune system to detect different types of microbial insult or danger. For example, TLR4 recognises lipopolysaccharide (LPS), thus enabling innate immune cells to specifically sense Gram-negative bacteria. Finally, the innate immune system had lost its 'non-specific' tag.

Innate immunity can be divided into two major components: cellular mediators and soluble mediators (Kapetanovic and Cavaillon 2007). Cellular mediators (monocytes, macrophages, granulocytes, mast cells, epithelial cells, innate lymphoid cells, etc.) can detect and respond to danger signals (Fig. 6.1). They do so via PRRs that can recognise PAMPS, their molecular structures being relatively conserved through evolution. PRR-mediated activation of these cells switches on specific antimicrobial pathways and regulates the expression of a suite of pro-inflammatory genes, such as those encoding cytokines and chemokines, secreted proteins that coordinate communication between the different actors of the immune system. These cytokines affect not only other immune cells, but also exert systemic effects. For example, cytokines/chemokines target the bone marrow (increasing hematopoiesis), the central nervous system (inducing fever) and the liver (eliciting production of acute phase proteins). Amongst the many soluble mediators of innate immunity, the complement system has been extensively studied. Although originally described by Jules Bordet in 1896, the name *complement* was first coined by Nobel Prize laureate, Paul Ehrlich, as this heat-sensitive serum factor could complement the action of antibodies. The complement system relies on a central element, the C3 convertase that, when activated, triggers a cascade of activation of different downstream proteins. The end result is the

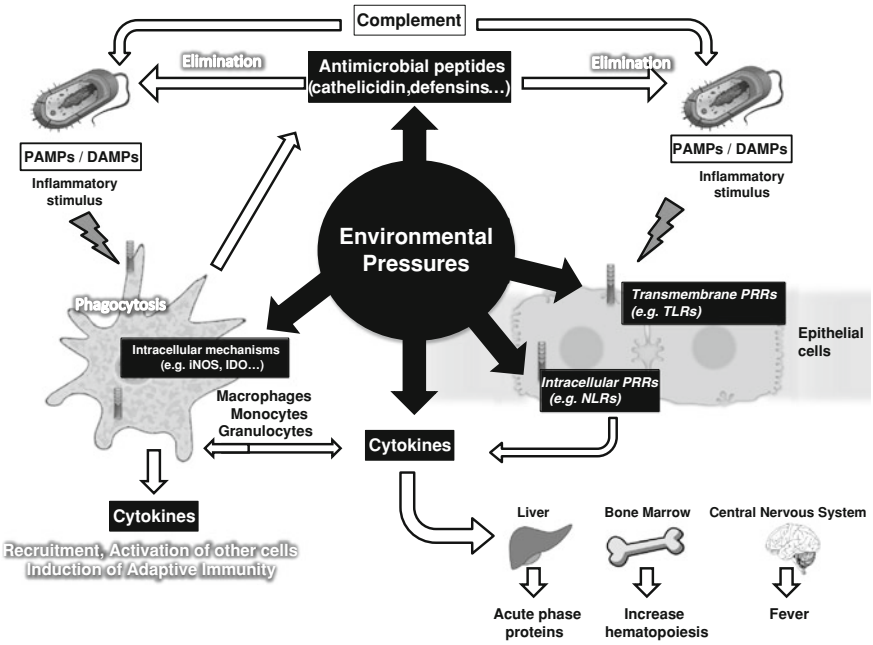


Fig. 6.1 *The different components of innate immunity and their susceptibility to environmental pressure*. Pathogens and their products, as well as endogenous danger signals, are detected by PRRs such as TLRs and NLRs. In response, epithelial cells and innate immune cells such as macrophages switch on inflammatory genes such as cytokines to communicate with other immune system compartments and also express antimicrobial products to directly eliminate microorganisms. As a result of evolutionary pressure from both pathogen and host danger signals, the genes encoding these innate immune mediators are susceptible to evolutionary change

generation of specific peptide mediators that can promote chemotaxis (e.g. C5a), opsonins that enhance microbial clearance by phagocytosis (e.g. C3b), and the membrane attack complex C5b6789n that directly lyses microorganisms by membrane disruption. Interestingly, complement is probably one of the most ancient of innate immune weapons in multicellular organisms, as it has been estimated that duplication of the different complement genes occurred around 500 million years ago (MYA) (Nonaka and Kimura 2006).

Innate immunity is often viewed as a biological system designed to provide the first line of defence against infectious diseases. In fact, it serves a much broader function than this, detecting danger in the form of not only infections, but also tissue damage and dysregulated cellular and metabolic functions. This concept, originally championed by Polly Matzinger (1994), then subsequently by Joost Oppenheim who coined the name 'alarmins' (Oppenheim and Yang 2005) and others, proposed that the immune system does not exclusively recognise 'non-self' signals, but more broadly responds to danger signals occurring during cell distress, damage, destruction and death (the 4 D's of the danger model). For example, the

host protein HMGB1 is an alarmin that is released as a response to the 4 D's to activate innate immunity. Thus, the innate immune system acts as a homeostatic rheostat, which is designed to detect perturbations and return the system to homeostasis. When it fails to do so, inflammation-related pathologies can result.

6.1.2 *Mechanisms that Drive Evolutionary Differences Between Species*

Model biological systems have provided powerful insights into innate immune processes, as evidenced by the initial description of phagocytosis by Mechnikov through his studies on starfish larvae (Tauber 2003), the identification by Lemaitre and Hoffmann of toll as a host defence gene in drosophila (Lemaitre et al. 1996), and Bruce Beutler's genetic mapping of TLR4 as the LPS receptor in mouse (Poltorak et al. 1998). However, conservation in biological pathways is essential if one is to using findings from model organisms to infer biological processes in humans. Although this is often the case, it is not always so; indeed, some biological systems are particularly susceptible to evolutionary change. This is the case with sensory systems, for example. Young et al. showed that mice possess three times the number of olfactory receptor genes to humans, as a result of both frequent duplications in the mouse genome and deletions in human (Young et al. 2002). Similarly, innate immunity is particularly susceptible to evolutionary change, due to strong selection pressure from rapidly evolving pathogens. The propensity of a system for evolutionary change is perhaps most readily apparent when examining intraspecies variability. Studies in pigs, for example, have shown that genes that vary most in their expression pattern across a population were clearly enriched for immune function, for example CXCL10, IDO1 and IL1RN (Dawson et al. 2013; Kapetanovic et al. 2013). This high variance in the expression of immune genes was also reported in both mouse and human. Murphy et al. studied 615 genes and found that the strongest divergence was apparent for the 75 genes involved in host defence, whereas genes encoding cell cycle regulators or structural proteins were highly conserved (Murphy 1993). More recently, dramatic inter-individual variation in innate immune response genes was demonstrated in studies of human monocytes responding to LPS (Fairfax et al. 2014).

Phenotypic differences between species can arise from multiple mechanisms. Gene duplication, described by Susumu Ohno in 1970 (Ohno 1970), is one of the major mechanisms driving evolutionary divergence. Duplicated genes (paralogues) can diverge and have different expression patterns and functions (Rensing 2014). Mutations are often deleterious (e.g. frameshift or the insertion of a stop codon) and cumulative, thus leading to non-functional pseudogenes. However, some paralogues acquire an entirely new function that was not present in the first place (neofunctionalisation). For example, the eosinophil cationic protein (ECP), an innate immune effector protein, diverged from a ribonuclease family and has undergone neofunctionalisation in primates (Zhang et al. 1998). The ancestral ribonuclease

proteins are also present in eosinophils and act as antiviral agents by degrading genomic RNA of retroviruses. By contrast, ECP has direct antimicrobial activity against parasites and bacteria (Lehrer et al. 1989). Interestingly, ECP has very low RNase activity (Young et al. 1986). Finally, duplication can lead to subfunctionalisation, a neutral mutation process dividing the function of the original gene between the two copies. This type of duplication offers more flexibility (Lynch and Conery 2000). A good example of this in innate immunity is the chemokine CXCL12 in zebrafish where two paralogues have complementary functions (Boldajipour et al. 2011).

Gene duplication can arise during cellular division through different mechanisms; unequal crossing-over during meiosis, replication errors by DNA polymerase or aneuploidy, although the latter is often embryonically lethal (Kaessmann 2010). Duplication can also occur through the actions of endogenous retroviruses, which are present in mammalian genomes. These viruses can trigger retrotransposition, where enzymes reverse transcribe RNA into cellular DNA, leading to a new copy of the gene. As a consequence, these retrogenes lack introns and have polyA tails. Since their promoter region is absent, these duplications are usually not expressed (Hurles 2004). Differences between species in retrotransposition can thus have profound effects on genome evolution. As one example, comparing human and mouse, the CDY-related (for chromodomain Y) gene family ancestor has been duplicated in all mammals into two autosomal genes CDYL and CDYL2, which encode enzymes with histone acetyltransferase activity. The human and mouse CDYL2 genes are highly conserved, but Dorus et al. (2003) found that, before human and mouse divergence, CDYL was retrotransposed on the Y chromosome. Interestingly, the human species has retained the copy, whereas the mouse lineage has lost this retrogene. CDY in human is only expressed in testis and may have given rise to new spermatogenic functions. A classic example of relevance to innate immunity is the tumour necrosis factor gene (TNF), which encodes a potent pro-inflammatory cytokine. Only one TNF superfamily (TNFSF) member is present in sea anemone and drosophila, but multiple rounds of duplication have led to 18 different TNFSF members in human, dispersed on four different chromosomes, including TNF, as well as the related family members lymphotoxin-alpha and lymphotoxin-beta (Wiens and Glenney 2011). Since TNFSF members play a role in regulating cells of the immune system, the extensive duplication of this gene family in vertebrates likely reflects the development of the adaptive immune system and a requirement for increased complexity in controlling this system.

In addition to the evolution of non-orthologous genes, strict one-to-one orthologues can also contribute to phenotypic differences between species. Amino acid changes can alter protein function, for example altering ligand specificity of cell surface receptors, but differences in gene expression patterns, which can occur through changes in gene regulatory sequences including promoters, enhancers and non-coding RNAs, can also have profound influences on biological functions of individual genes. Numerous studies have demonstrated high divergence in gene regulatory regions between mouse and human. This is the case for the

apolipoprotein E (APOE) gene (Maloney et al. 2007), which is differentially regulated in human versus mouse macrophages (Irvine et al. 2009), as well as the P-selectin gene (Pan et al. 1998), which mediates cell:cell adhesion to enable leukocyte migration to sites of inflammation. In the case of P-selectin, the mouse promoter contains tandem GATA and AP1-like sequences that are absent from the human promoter. Such variations within promoter regions have been of particular interest for some of our own studies (see Sect. 4.3). The above mechanisms that drive evolutionary divergence can also be accelerated by the different generation times of individual species. For example, mice have a much shorter generation time than humans, so have a much greater window of opportunity to be affected by such mechanisms, as compared to humans. As a result, the number of gene duplication events is much higher in mice than in human. Studies from Church et al. (2009) estimated the number of gene duplication events to be 3,767 in the mouse, compared to 2,941 in human. Given that mice and humans evolved from a common ancestor ~70–80 MYA (Waterston et al. 2002), the impact of generation time on evolutionary differences is not an insignificant consideration.

Below, we discuss some examples of innate immune differences between humans and mice at the genetic level, as well as the mechanisms responsible. We then provide more in-depth discussion around differences between human and mouse macrophages in their responses to LPS, again with a focus on genetic mechanisms.

6.2 The Mouse as a Model to Study Human Innate Immunity

6.2.1 The Utility of the Mouse as a Model for Understanding Human Biology

Mice were first used as model organisms for genetic studies in the early 1900s. The French biologist, Lucien Cuénot set the foundation for mouse genetic studies by demonstrating that genetic inheritance, previously observed by Mendel in his studies on plants, applies also to mammals, as assessed by inheritance of coat colours (Gayon and Burian 2000). In 1903, Clarence Cook Little, as a student of William Castle, generated the first inbred strain of laboratory mice carrying recessive genes Dilute, Brown, non-Agouti, today known as the DBA strain (HD 1918; J 1966). This strain was followed by the creation of the now widely used strains C3H, C57BL and BALBc. Nobel Prize laureate, George Snell, another student of William Castle, also contributed extensively to the field of mouse genetics by establishing congenic strains and discovering the genes controlling the major histocompatibility complex (MHC) (Klein 2001). Using rodents to understand the human immune system is, at the first glance, particularly convenient. They are mammals that share very similar embryonic development patterns and

mechanisms, physiology, behaviour and can be used to model many human diseases. Their use as a model system is also facilitated by their large litter sizes and by the fact that they become sexually mature at only 7–8 weeks after birth. Moreover, the initial sequencing of the human and mouse genomes demonstrated that over 95 % of the mouse genome is similar to the human one, and first genomic comparisons estimated that only 300 genes were unique to each species (Waterston et al. 2002). More recent studies have estimated that the mouse has 20,210 protein-coding genes and that 75 % of mouse genes have a 1:1 orthologue in human (Church et al. 2009). Shay et al. (2013) reported similar results, and suggested that only a minority of orthologous genes (169) showed differing expression patterns between human and mouse.

The widespread use of mice for many decades has also driven the development of many molecular and biotechnological tools, including antibodies, recombinant proteins and gene-modifying technologies, for studies in this species. In the late 1980s, the development of procedures to manipulate mouse embryonic stem (ES) cells led to the ability to silence specific genes in mice (Goldstein 2001). Martin Evans's team at Cambridge University demonstrated that ES cells derived from mouse blastocysts could be isolated in vitro, genetically altered and inserted into a new embryo. This discovery was strengthened by another technology developed by Mario Capecchi and Oliver Smithies that allowed researchers to precisely target single genes by homologous recombination, and led to the creation of the first gene targeted (knock-out) mouse in 1989 (Koller et al. 1989). These two combined technologies opened a completely new path in molecular biology for which these investigators were awarded the Nobel Prize in physiology or medicine in 2007 for their discoveries of 'principles for introducing specific gene modifications in mice by the use of embryonic stem cells'. Today, this technology has been adopted on a genome-wide scale for gene targeting, thus allowing researchers to study the role of one specific gene on organismal development, physiology and pathophysiology. More recently, tools for genetic manipulation have been further improved and very specific modifications (e.g. changing specific amino acids or adding epitope tags) can now be rapidly engineered. For example, CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats) is a bacterial defence system, which has recently been developed as a gene editing technology for gene modification in various species, including mice (Jinek et al. 2012). Similarly, TALEN is another recently described method which uses a nuclease from *Xanthomonas* bacteria for gene editing (Zhang et al. 2011). Numerous mouse models of infection and inflammatory diseases have been developed, many of which have been successfully applied to understanding or treating human diseases (Webb 2014). For example, tumour-necrosis factor (TNF) was shown to play a major role in chronic inflammation in models such as mouse collagen-induced arthritis, and subsequently, targeting TNF with antibodies or soluble receptors has been translated into effective therapies for rheumatoid arthritis and inflammatory bowel disease (O'Dell 2004; Kollias et al. 2011). Since the development of technologies to genetically manipulate mice for in vivo studies, researchers have largely used the

mouse to model the human immune system, but nonetheless, it is clear that the immune systems of these two species show many evolutionary differences; as with any model, the mouse as a model to study human immunity has some deficiencies.

6.2.2 Limitations of the Mouse Model

As alluded to above, some biological systems are particularly susceptible to evolutionary change. This is the case for genes involved in reproduction, perception and immune responses (Church et al. 2009). Consequently, it is inevitable that some differences in immune responses between humans and mice will impact on the utility of the mouse as a model for studying human immune system biology. There are numerous phenotypic and genotypic differences between humans and mice; these have been well described elsewhere (Mestas and Hughes 2004). Genetic differences not only reflect differences in gene numbers, but also in the way that orthologous genes are expressed or function. Below, we briefly provide an overview of some of the key phenotypic differences (e.g. physiology), before providing a more detailed analysis of genetic differences in innate immunity.

Mice are ~3,000 times smaller than humans and have a slower basal metabolic rate (Demetrius 2005). Blood circulation also varies between the two species. The mouse heart beats very quickly (400–600 beats per minute, bpm) compared with humans (60–90 bpm). The presence of specific immune organs, such as the bronchus-associated lymphoid tissue (BALT), can also differ between species (Pabst and Gehrke 1990). Other organs such as the eye and the dermis (Lei and Yao 2006; Henriksson et al. 2009), as well as the process of wound healing (Wong et al. 2011), also show some differences between human and mouse. At a cellular level, the white blood cell compartment in humans is neutrophil-rich (50–70% neutrophils), whereas in mice it is lymphocyte-rich (75–90% lymphocytes) (Doeing et al. 2003). Specific cell types in mouse versus human can also express different markers, which can influence responsiveness to the extracellular environment (e.g. growth factors or pathogens). For example, the cytokine receptor flt3 is a key marker of hematopoietic stem cells in human, but not mouse (Sitnicka et al. 2003). Similarly, Crocker et al. (1987) demonstrated that CD4, a receptor for HIV expressed by a subset of T cells, is also expressed on macrophages from human, but not mouse.

Such physiological and cellular differences can sometimes cause unforeseeable problems for the translation from mouse models of immune-related studies to human clinical trials. This is the case for some clinical trials using antibodies targeting components of the immune system. As an example, interferon-gamma (IFNγ) has a protective effect in demyelinating disease in a mouse model, but human trials were stopped as the treatment actually exacerbated multiple sclerosis in patients (Panitch et al. 1987). Many problems in translating studies from mice to the clinic stem from inappropriateness of the model used. For example, human immune-related diseases often target the young, who are particularly susceptible to

infectious disease or the elderly who suffer from infectious, metabolic and inflammatory disease due to a decline in immune function. However, immune studies in mice are typically performed on young adults (6–8 weeks of age), with very few studies exploring immune-related diseases in (more relevant) young and aged mice. Nonetheless, evolutionary differences between mice and humans are undoubtedly a contributing factor as well, and understanding genetic differences between the murine and human immune systems is thus important for translation of mouse immune studies to the clinic.

6.3 Genetic Differences in Innate Immunity

Several families of innate immune genes exhibit differences between human and mouse in their repertoire and/or in the functions of one-to-one orthologues. These include genes encoding secreted molecules such as antimicrobial peptides (e.g. defensins, cathelicidins), enzymes (e.g. lysosyme) and cytokines and chemokines (e.g. IL-8/CXCL8). Such differences are also apparent for many genes encoding transmembrane proteins (e.g. TLRs, NLRs, cytokine receptors) that transduce signals to innate immune cells to enable appropriate cellular innate immune functions. Below, we discuss specific examples of such differences, before providing a more detailed overview of LPS signalling through TLR4, as a case study of human versus mouse differences in innate immunity.

6.3.1 Defensins and Other Mediators with Direct Antimicrobial Functions

Defensins are probably one of the oldest host defence systems, as they are found in both animals and plants (Thomma et al. 2002). Indeed, the *Arabidopsis thaliana* genome encodes 11 different plant defensins. These cationic polypeptides are usually small (between 18 and 45 amino acids), cysteine-rich and are generated from a pro-protein by enzymatic cleavage. Their primary function is to bind to microbial cell membranes and form pores leading to microbial lysis, but they also have a range of other immune-modulating properties. Three classes, the alpha (α), beta (β) and theta (θ) defensins, are distinguished by the position of cysteine residues and by their disulphide bonding patterns. All three classes are present in mammals, but there are several human versus mouse differences within these families. Humans possess 6 α- and around 40 β-defensin genes or pseudogenes (Lehrer and Lu 2012; Semple and Dorin 2012). Human neutrophils express four of the α-defensins (human neutrophil peptide [HNP] 1–4), while two (HD-5 and HD-6) are expressed by intestinal paneth cells that are localised in the crypts of the small intestine. The first 4 α-defensins were reported to be absent from mouse neutrophils at the protein level (Eisenhauer and Lehrer 1992). Early studies also

reported that paneth cells in mice express transcripts for 20 different alpha-defensins (cryptidins), whereas only two of these are found in human (Ouellette and Selsted 1996). Subsequent studies have confirmed the rapid evolution of many α-defensin genes, with even different inbred mouse strains showing distinct repertoires of these genes (Shanahan et al. 2011). The maturation process for paneth cell defensins is also reported to be different between the two species. Defensin-5 is stored as a pro-form in secretory vesicles and is processed by matrix metalloproteinase (MMP) 7 in mice, whereas it is processed by trypsin in humans (Cunliffe et al. 2001; Ghosh et al. 2002). β-defensins are defined by a six-cysteine motif and are encoded by two exons. Although close to 40 open reading frames for β-defensins have been found, only six proteins, and an additional 11 RNA transcripts, have been confirmed in vivo. The mouse genome seems to have a larger number of β-defensin genes, although whether these genes all produce functional transcripts is yet to be confirmed (Schutte et al. 2002). θ-defensins exist as functional defensins in mouse, but as six pseudogenes in human, as they are terminated by a premature stop codon (Lehrer et al. 2012). The loss of these defensins appears to have occurred after divergence with orangutan (~7.5 MYA). Interestingly, recent research to 'reactivate' θ-defensins by using aminoglycosides to read through the premature stop codon resulted in human myeloid cells producing cyclic antiviral retrocyclins that inhibit HIV-1 infection (Venkataraman et al. 2009). Retrocyclin-1 inhibits the production of proviral DNA, suggesting that it prevents viral entry (Munk et al. 2003). It is interesting to speculate as to whether the loss of θ-defensins was due to reduced environmental pressure of viruses in humans or whether they were made redundant because of the duplication of α and β defensins. Interestingly, the α-defensin HD-6/DEFA6 is not directly antimicrobial, but it was recently shown to form fibrils and nanonets that aggregate bacteria and prevent their attachment and subsequent invasion of epithelial cells (Chu et al. 2012). Genetic deficiency in this defensin, which has no orthologue in mice, was also linked to the development of Crohn's disease in humans (Wehkamp et al. 2009). Additionally, transgenic mice expressing human α-defensin HD-5/DEFA5, which is also absent from mice, were protected against enteric salmonellosis (Salzman et al. 2003). These observations therefore highlight key roles in humans for specific α-defensins in gastrointestinal host defence and inflammation, which studies with traditional mouse models are unable to capture. Finally, concerning the β-defensins, while the first exon that encodes the signal sequence is relatively well conserved, the second one that encodes the mature peptide region shows a low level of similarity across species and evidence of positive selection (Maxwell et al. 2003). These reports, and others, suggest that the defensin family is under strong evolutionary pressure, with gene duplication resulting in specialised roles for individual defensins.

Cathelicidins, another family of mammalian antimicrobial peptides, are expressed by macrophages, neutrophils and other cell types. Although they are heterogeneous in amino acid sequence, size and structure across species, they all contain a relatively conserved N-terminal region known as the cathelin domain, which is cleaved to release the active antimicrobial peptide (Zanetti 2004). Many

form α-helices spanning approximately 23–37 amino acids. This is true for the sole mouse cathelicidin (cathelin-related antimicrobial peptide, CRAMP) and the sole human cathelicidin LL-37. Despite the one-to-one orthology relationship for the human and mouse cathelicidins, there are some differences in their regulation. LL-37 is TLR-inducible via a vitamin D-dependent pathway in human macrophages, while CRAMP is not (Liu et al. 2006). Interestingly, overexpression of the human LL-37 in mouse increased their resistance to bacterial challenge, suggesting that this molecule may possibly have additional functions to mouse CRAMP in host defence pathways (Bals et al. 1999).

Lysozyme, or muramidase, is a hydrolase that can damage bacterial cell walls by catalysing the hydrolysis of the cell wall component peptidoglycan. It was discovered by Alexander Fleming in the 1920s, where it was described as a molecule contained in secretions such as saliva or mucus that helps protect the epithelial cell layer (Fleming 1922). Lysozyme is present in the granules of neutrophils, but is also a major secretory product of macrophages (Gordon et al. 1974). Humans have only one lysozyme gene, but mice possess two, lysozyme M and P. The duplication of the mouse lysozyme gene occurred ∼5–10 MYA, after mouse and rat diverged (Cortopassi and Wilson 1990).

Human versus mouse differences in antimicrobial effectors are not only confined to secretory products. Members of the interferon-inducible GTPase family reside on intracellular membranes and relocalise to vacuoles after infection or phagocytosis, where they are thought to act directly on pathogens to exert microbicidal effects (Martens et al. 2004). This family includes four groups of proteins: the myxovirus resistant proteins (Mxs), the guanylate-binding proteins (GBPs), the very large inducible GTPase proteins (VLIGs) and the p47 immunity-related GTPase proteins (IRGs) (Li et al. 2009). This last gene family shows substantial divergence between human and mouse. There are 23 IRG genes in mouse divided into five subfamilies (IRGA, IRGB, IRGC, IRGD and IRGM), yet only two functional human paralogues (IRGC and IRGM), along with an additional truncated pseudogene (IRGQ) (Bekpen et al. 2005; Martinez et al. 2006). IRGM is the only human IRG likely to play a role in immunity. It has been linked to autophagy and the elimination of *Mycobacterium tuberculosis* (Singh et al. 2006). The second human IRG, IRGC, is highly conserved, not regulated by cytokines and is expressed in testis, which would suggest that it does not play any role in immunity. IRGs are expressed in other mammals, but without any apparent phylogenetic order. For example, dogs possess six IRGs, while cats possess none (Gazzinelli et al. 2014). The expression of IRG family members in mouse offers resistance against *Toxoplasma gondii* (Gazzinelli et al. 2014). Its action results in the rupture of the vacuole membrane containing the parasite and allows for its degradation in the host cell cytoplasm (Howard et al. 2011). Indeed, mice that lack Irgm1 or Irgm3 become highly susceptible to this parasite. Host-pathogen co-evolution represents a constant balance between new antimicrobial defences from the immune cells and new evasion systems of the pathogens. Illustrating this, some parasites can secrete a pseudokinase (ROP5) and a kinase (ROP18) that can, in combination, inactivate IRG proteins, resulting in high intracellular survival

(Hunter and Sibley 2012). In response, some cells in mice can express a 'tandem' of IRG proteins that can act as a decoy for the pseudokinase complex (Lilue et al. 2013). Similar to IRG, the p65 guanylate-binding protein 1 participates in the host antiparasite response via autophagy (a mechanism that degrades intracellular components) in combination with different proteins in human (ATG16L1) and mouse (Atg7 and Atg16L1). Recent work suggests that the IFNγ-inducible IRG family member GBP1 may contribute to host defence by inducing actin remodelling around parasite-containing vesicles (Ostler et al. 2014).

6.3.2 PRRs: TLRs

As previously alluded to, the role of the toll protein in drosophila host defence was discovered by Jules Hoffman in 1996, while TLR4 was identified the following year as the receptor for endotoxin/LPS in humans (Lemaitre et al. 1996; Medzhitov et al. 1997). After the initial discovery of TLR4, several other TLRs and their ligands were quickly identified. The human genome encodes 10 TLRs, whereas mice possess 12 TLRs. Interestingly, several simple multicellular organisms, which do not possess an adaptive immune system, have a greatly expanded repertoire of TLRs. For example, the sea urchin genome contains 222 different TLRs (Rast et al. 2006). Ligands for the majority of human and mouse TLRs have now been described. TLR2 can heterodimerise with TLR6 or TLR1 to recognise diacylated and triacylated lipoproteins, respectively (Takeuchi et al. 2001). Flagellin, the protein component of flagella that is involved in bacterial locomotion is sensed by TLR5 (Hayashi et al. 2001). A subfamily of TLRs detects nucleic acids. TLR9 recognises unmethylated CpG motifs found in bacterial and viral DNA (Bauer et al. 2001), and TLR3 recognises double-strand RNA, which is formed during viral replication (Alexopoulou et al. 2001). TLR7 and TLR8 can detect synthetic compounds of the imidazoquinoline family, as well as single-strand RNA from specific viruses (Heil et al. 2004). For all of the TLRs described above, there are clear one-to-one orthologues in human and mouse. In contrast, TLR10 is a pseudogene in mouse, whereas TLR10 is functional in human; it encodes a receptor that can heterodimerise with TLR2 or TLR1 (Hasan et al. 2005). Human TLR10 binds triacylated lipopeptides and shares overlapping ligand specificity with TLR1, although it elicits fundamentally different signalling outcomes (Guan et al. 2010). More recent studies suggest roles for TLR10 in antiviral responses (Lee et al. 2014) and in responses to *Listeria monocytogenes* (Regan et al. 2013). Unlike TLR10, TLR11 and 12 are present in mouse, but not human. These TLRs sense uropathogenic bacteria, as well as a profilin-like protein from the parasite *Toxoplasma gondii* (Yarovinsky et al. 2005; Koblansky et al. 2013). Finally, TLR13, which is also absent from human, detects a specific sequence within bacterial 23S ribosomal RNA (rRNA) (Bordon 2012; Oldenburg et al. 2012). Thus, there are substantial differences in the repertoire of TLRs between humans and mice, resulting in differences in either the specific PAMPs that can be

sensed by human versus mouse or in the specific mechanisms that mediate sensing of individual PAMPs. However, there can also be differences in the ligand specificity for individual TLRs with one-to-one orthology relationships. This is not surprising, given that evolutionary divergence between TLRs is mostly apparent in the extracellular domains, as compared to the cytoplasmic domains (Zhou et al. 2007). For example, in mouse, both TLR2/6 and TLR2/1 heterodimers can recognise *Aspergillus fumigatus*, whereas, in human, only the latter TLR complex can detect this pathogen (Rubino et al. 2012). In addition, mouse TLR8, unlike human TLR8, does not recognise imidazoquinoline compounds (Jurk et al. 2002). Consequently, this receptor was first thought to be non-functional, although a later study has suggested that murine TLR8 in plasmacytoid dendritic cells detects vaccinia virus DNA (Martinez et al. 2010). A more recent study has also shown that TLR8 constrains the function of TLR7 in mice, to prevent pathological inflammatory responses (Desnues et al. 2014). Differences in regulatory elements can also result in differences in the expression patterns of the TLR genes between species. For example, TLR2 is generally weakly expressed by peripheral blood leukocytes in mice, but highly expressed in the spleen, lung and thymus (mostly T cells). Its expression can be further upregulated in many cells, including T cells, upon activation (Matsuguchi et al. 2000). In contrast, TLR2 is constitutively expressed in human, although it is still somewhat inducible (Visintin et al. 2001). Similarly, expression of both TLR3 (Heinz et al. 2003) and TLR6 (Schroder et al. 2012) is LPS-inducible in macrophages from mouse, but not human. In the case of TLR3, the difference in human versus mouse LPS-regulated expression correlated with divergence in promoter architecture and sequence; whereas the human TLR3 promoter is a TATA-containing promoter, the mouse promoter is not (Rehli 2002; Heinz et al. 2003).

The susceptibility of TLRs to evolutionary change is also apparent when considering intraspecies variation. Barreiro et al. (2009) studied differences in TLRs during recent human evolution. Through comparisons of African, European and East-Asian populations, they found that intracellular TLRs, which are particularly important for viral recognition, have been under strong purifying selection, meaning less variability. Individuals presenting TLR7, TLR8 or TLR9 deficiencies have never been reported, suggesting that these receptors play non-redundant roles in host defence. In contrast, cell surface TLRs have tolerated higher rates of deleterious mutations, thus suggesting higher immunological redundancy for these receptors. Within the cell surface TLRs, the TLR1–TLR6–TLR10 cluster, which is located in the same region at chromosome 4p14, shows clear evidence of positive selection in European and East-Asian populations, as compared to African populations. This may reflect the evolutionary shaping of these receptors in response to emerging pathogens. There is certainly evidence that this variation across this TLR cluster may have been shaped by co-evolution with *Yersinia pestis*, for example (Laayouni et al. 2014). Interestingly, this subfamily of lipopeptide-sensing TLRs appears to have been subjected to more species-specific adaptation than other TLRs. This includes the orientation of these genes within the

cluster and the number of paralogues. For example, fish possess only one TLR1 gene, most mammals, except for opossum, have three paralogues (TLR1-6-10), while chicken has only two (Roach et al. 2005).

6.3.3 PRRs: NLRs

As transmembrane receptors, the TLRs detect PAMPs present in the extracellular compartment or in endosomes. Innate immune cells also possess PRRs that detect danger signals within the cytoplasm, thus enabling detection of pathogens within different intracellular locations. The most widely studied of the cytoplasmic danger sensing molecules are the NLRs, which typically contain a leucine rich repeat-containing (LRR) domain, a nucleotide-binding domain and either a caspase-activation and recruitment domain (CARD) or a PYRIN domain. NLRs have been classified according to domain architecture (NLRA, NLRB, NLRC1-5, NLRP1-14 and NLRX1) (Ting et al. 2008) or on the basis of phylogenetic analysis (NLRP, IPAF and NOD subfamilies) (Schroder and Tschopp 2010). NLRs seem to have appeared before the TLRs and are homologous to proteins encoded by R (resistance) genes in plants (DeYoung and Innes 2006).

Within the 22 human NLRs and more than 30 mouse NLRs (Kanneganti et al. 2007), two well-known orthologous (1:1) receptors are NOD1 and NOD2, also known as NLRC1 and 2, which drive NF-κB-dependent transcriptional responses to specific pathogen products (Philpott et al. 2014). NOD1 detects y-D-glutamyl-meso-diaminopimelic (iE-DAP), a specific dipeptide present in peptidoglycan from Gram-negative bacteria (Chamaillard et al. 2003; Girardin et al. 2003a). The NOD2 ligand, muramyldipeptide (MDP), is present on both Gram-positive and Gram-negative bacteria (Girardin et al. 2003b). Interestingly, the macrophage-activating properties of MDP was discovered in 1978 by Louis Chedid's team (Specter et al. 1978), but it took another 25 years to finally identify the receptor responsible for these effects. Although there are substantial differences in the NLR repertoire between mouse and human, NOD1 and NOD2 have a clear 1:1 orthology relationship. Nonetheless, some species differences have been reported; in mouse, but not human, NOD1 responds to tracheal cytotoxin, a toxin released from *Bordetella pertussis* (Magalhaes et al. 2005). This study also identified differences between human and mouse NOD1 in the specific peptidoglycan product that they each recognise. Also, NOD1 expression was reported to be strongly up-regulated in activated murine monocytes, in contrast to human monocytes (Lech et al. 2010).

A subset of NLRs are capable of forming a cytoplasmic signalling platform called the inflammasome (Schroder and Tschopp 2010). Upon ligand recognition, most inflammasome-forming NLRs promote the clustering of the ASC adaptor molecule into a signalling hub (often termed 'the ASC speck'), which recruits the pro-inflammatory caspase-1 and one or two HIN-200 DNA. Activated caspase-1 is able to cleave pro-forms of certain inflammatory cytokines (e.g. IL-1β, IL-18) to enable their maturation and subsequent release, and also initiates a form of inflammatory

cell death called pyroptosis. Several poorly characterised NLRs that are likely to form inflammasomes show human:mouse differences. For example, human NLRP7, NLRP8, NLRP11 and NLRP13 do not have murine orthologues because of either primate-specific evolution (NLRP7, 11) or loss during rodent evolution (NLRP8, 13) (Ariffin and Sweet 2013). Even well-characterised inflammasome-forming NLRs (NLRC4, NLRP1 and NLRP3) that have clear human and mouse orthologues can exhibit differences between these species in either their expression or their activation pathways. For instance, human NLRP3 recognises the Gram-negative bacterial pathogen *Francisella tularensis*, whereas mouse NLRP3 does not (Atianand et al. 2011). Nevertheless, mouse macrophages can still detect this pathogen, relying on another inflammasome-forming PRR, AIM2 (Fernandes-Alnemri et al. 2010), which is described below. AIM2 also senses this pathogen in human cells. Thus, humans have an additional mechanism for inflammasome activation in response to this pathogen compared to mice. NLRP10, the only NLRP family member lacking an LRR, is also reported to display some mechanistic differences between human and mouse. Human NLRP10 was described to inhibit procaspase-1, the production of IL-1β and ASC speck formation, whereas mouse NLRP10 had a similar functional role, but did not inhibit ASC aggregation (Imamura et al. 2010). A recent study from Su et al. (2013) has also reported that the structure of human NLRP10 is distinct from its mouse counterpart. This work shows that the helix H3 and loop H2-H3 adopt a specific conformation, which is not observed in mouse, and could explain species differences in the interaction with ASC. Another report documented a role for mouse NLRP10 in the initiation of adaptive immunity in dendritic cells, rather than in negative regulation of the inflammasome (Eisenbarth et al. 2012), although this potential function has not yet been studied in human cells. Interestingly, a deletion within exon 2 of human NLRP10, which occurred since the divergence of human and chimpanzee, results in the loss of 30 amino acids from the C-terminus and also generates an alternative 3′-UTR (Ha et al. 2009). This could also potentially contribute to differences in function and/or regulation of human NLRP10. There is also evidence for functional divergence between human and mouse in downstream inflammasome signalling pathways. Although caspase-1 was originally presumed to be the only caspase downstream of inflammasome activation, roles for other pro-inflammatory caspases are emerging. One of these is caspase-11 that, in mouse, is involved in sensing cytoplasmic LPS (Hagar and Miao 2014). Caspase-11 does not have a strict 1:1 orthologue in human, as it is similarly related to both caspase-4 and caspase-5. This could result in enhanced complexity in inflammasome signalling in human cells, if indeed both of these caspases are involved in this pathway. Caspase-8, a pro-apoptotic caspase, has also been shown to be a downstream target of inflammasome activation (Antonopoulos et al. 2013; Shenderov et al. 2014; Gurung et al. 2014). Caspase-8 deficiency is embryonically lethal in mice, whereas humans can survive without this gene, albeit with immunodeficiency (Chun et al. 2002). This suggests that there are differences in caspase-8 function in human versus mouse, although whether any of these differences relate to inflammasome responses is unknown.

6.3.4 PRRs: The RIG-I-Like Receptors

In addition to the NLRs, another family of intracellular receptors also act as cytosolic PRRs, but contain a helicase domain, rather than an LRR. This family includes the retinoic-acid-inducible gene I (RIG-I), the melanoma differentiation-associated gene 5 (MDA-5) and the regulator laboratory of genetics and physiology 2 (LGP2). These three molecules are present in both human and mice and their activation by viral nucleic acid leads to the induction of pro-inflammatory cytokines and type I interferon (IFN) (Eisenacher and Krug 2012). A study of 186 humans from different origins showed that RIG-I has a low level of diversity, suggesting that this gene is under strong evolutionary constraint. Conversely, evidence of positive selection for non-synonymous variants of both MDA-5 and LGP2 has been reported (Vasseur et al. 2011). The low tolerance of RIG-I to evolutionary change might reflect the importance of this receptor in recognition of viral RNA, but could also indicate an essential developmental role for RIG-I since deficiency in this gene is embryonically lethal in mouse (Kato et al. 2005). Some species-specific characteristics have been described for RIG-I-Like Receptors (RLRs). Lech et al. (2010) showed that, although the levels of RIG-I and MDA-5 were similar in resting human and mouse monocytes, stimulation strongly up-regulated these genes in mouse, in contrast to human monocytes.

6.3.5 PRRs: DNA Sensors

PyHIN proteins, which are encoded by four genes in humans (*IFI16*, *MNDA, AIM2* and *IFIX*) and 13–14 in C57BL/6 mice, are characterised by an N-terminal PYD and one or two HIN-200 DNA-binding domains at the C-terminus (Cridland et al. 2012). Based on phylogenetic analysis of both pyrin and HIN domains, AIM2, which is distinct from the NLRP family but forms an inflammasome upon detection of cytosolic DNA (Warren et al. 2010b), is the only PyHIN gene with a direct orthologue in the human and mouse genomes. IFI16 and IFIX reportedly arose from gene duplications in primates, and other PyHINs show domain-based speciation (Cridland et al. 2012). Comparative expression and sequence analysis of human and mouse PyHIN genes did not yield any similarities, with the exception of AIM2, supporting the theory that PyHIN family members have undergone species-specific duplication and divergence (Brunette et al. 2012). Nevertheless, IFI16 was reported to possess functional homology with murine p204, with which it shares a similar domain structure, but only 37 % amino acid identity. Both IFI16 and p204 were shown to act as cytoplasmic DNA sensors that induce IFN-α/β and other pro-inflammatory mediators (Unterholzner et al. 2010). On the other hand, murine p202, which acts as a negative regulator of AIM2-dependent inflammasome activation (Yin et al. 2013), has no known human homolog (Ru et al. 2013). It is possible that a functional equivalent of p202 exists in the human genome, despite the lack of a direct orthologue. Evidence for this has not yet emerged.

6.3.6 PRRs: C-type Lectin Receptors

C-type Lectin Receptors (CLRs), which recognise carbohydrates such as β-glucans, mannose or fucose, constitute a superfamily of more than 1,000 proteins that can be broadly partitioned into three subgroups: soluble CLRs (e.g. mannose-binding lectin or MBL), type I transmembrane CLRs (e.g. DEC-205) and type II transmembrane CLRs (e.g. Dectin-1) (Kingeter and Lin 2012). Of these, the type II transmembrane CLRs have been most extensively studied in the context of pathogen recognition (i.e. as PRRs). These CLRs initiate signalling through the Syk tyrosine kinase and the CARD9 adaptor protein, leading to activation of NF-κB, pro-inflammatory gene expression and an appropriate inflammatory response. CLRs are widely distributed within the species and can even be found in invertebrates. Studies on the sea anemone, *Nematostella vectensis,* have predicted 67 C-type lectin genes (Wood-Charlson and Weis 2009). Sattler et al. (2010) have found that, despite some inversion due to Alu sequences, the orientation of the genes between primate and the rodent cluster are well conserved. Furthermore, the analysis of individual genes revealed a high level of sequence conservation, again suggestive of evolutionary constraint. Members of the DECTIN-1 gene cluster, located in the NK gene complex, also show high homology between themselves, as well as between species (Sattler et al. 2012). One homologue of human CLEC-1 has even been detected in an organism as evolutionary different as *Caenorhabditis elegans*, albeit with only 12 % similarity.

Some CLRs have a more complex human:mouse orthology relationship; for example, human BDCA-2 (CD303) and human dendritic cell (DC) immunoreceptor (hDCIR) have two (mDCAR and mDCAR1) and four (mDCIR 1-4) murine orthologues, respectively (Kanazawa et al. 2004). Other human CLRs such as DC-associated lectin-1 (DCAL-1) have no clear murine orthologue identified to date. Some CLR orthologues are also differentially expressed in humans versus mice (Lech et al. 2012). These differences can be tissue specific, for example GALEC1 is more strongly expressed in human muscle compared to mouse, but down-regulated in human kidney compared to mouse. Human DCAL-2 (MICL, CLEC12A) is detected on monocytes, neutrophils, eosinophils and basophils, while murine DCAL-2 is additionally detected on thioglycollate-elicited neutrophils and macrophages, as well as bone-marrow-derived dendritic cells (Marshall et al. 2004). The regulation of these genes can also be species-specific. Unlike human DC-SIGN, expression of the murine homologue (SIGNR3) was not up-regulated by IL-4 and IL-13 (Tanne et al. 2009). Such differences in CLR expression patterns may possibly reflect different roles in human versus mouse, and there is at least some evidence of this for some of the CLRs that show differential regulation in human versus mouse. For example, murine DC-SIGN is reportedly not involved in T cell–dendritic cell interactions and does not bind to some pathogens such as *Leishmania mexicana,* cytomegalovirus and HIV, which do interact with the human receptor (Colmenares et al. 2002; Kwon et al. 2002). Additionally, human DC-SIGN seems to have also evolved to recognise *Mycobacterium tuberculosis*

(Schaefer et al. 2008). Thus, individual CLRs can display highly specific and distinct roles in pathogen recognition and immune regulation in human and mouse, emphasising the need for careful evaluation, comparison and interpretation of functional information about innate immune roles for individual CLRs.

One final example of divergence between species, in the context of carbohydrate recognition, is the sialic acid Neu5GC. Its expression is widespread in mice and other mammals, but undetectable in humans (Varki 2001). The absence of Neu5GC is due to a frameshift mutation in the gene encoding the enzyme cytidine-monophosphate-N-acetylneuraminic acid hydroxylase. This occurred after the divergence between human and great apes (Chou et al. 1998). As a consequence, the level of Neu5Ac (the precursor of Neu5Gc) in human is much higher than in other animals (Brinkman-Van der Linden et al. 2000). Interestingly, human Siglec-9, a lectin that recognises sialic acid, binds to both Neu5Ac and Neu5GC. In contrast, Siglec-9 from other primates binds preferentially to Neu5GC (Sonnenburg et al. 2004). This suggests that human Siglec-9 has adapted to this change in available sialic acid ligands. While this would suggest that Siglec-9 functions by interacting with host rather than pathogen products, the loss of Neu5GC from humans may have affected susceptibility to specific pathogens. For example, increased levels of Neu5Ac in human could actually contribute to increased susceptibility to *Plasmodium falciparum*, since this parasite binds Neu5Ac on human erythrocytes to enable cellular invasion (Orlandi et al. 1992).

6.3.7 Cytokines and Their Receptors

Activation of PRRs such as TLRs, NLRs or DNA sensors triggers intracellular signalling cascades, leading to the production of a plethora of inflammatory mediators, including numerous cytokines (encompassing interleukins, interferons and chemokines) (Moser et al. 2004). Below, we very briefly consider some human:mouse differences for these secreted molecules and their receptors.

Up to 50 chemokines have now been identified in humans (Griffith et al. 2014), and these are critical for intercellular communication and cellular recruitment to sites of inflammation. Chemokines are classified into four subfamilies depending on the positioning of cysteine residues within the primary amino acid sequence: CC and CXC chemokines (the two major subfamilies), as well as the C and CXXXC chemokines (Nomiyama et al. 2001). Chemokines are typically clustered on specific chromosomal regions, indicative of rapid gene duplication. Each cluster of chemokines often have a specific cellular target, as is evident with the human CXC chemokines (4q12–13) that primarily act on neutrophils and the CC chemokines (cluster 17q11.2) that mainly target monocytes and lymphocytes. This may provide functional redundancy within a cluster, in keeping with the fundamental roles that these molecules play in inflammation (Zlotnik and Yoshie 2000).

The extensive duplication of chemokine genes has resulted in numerous examples of divergence between human and mouse in specific chemokines or their receptors. CXCL8/IL-8, which is a potent neutrophil chemoattractant, is present in human, but is absent from mouse. Mice do possess other chemokines such as CXCL1/KC, CXCL2/MIP-2 and CXCL5/LIX (Rovai et al. 1998) that are likely to fulfil similar roles to human IL-8 since these chemokines act on the same receptors as human IL-8, e.g. CXCR1 (Fan et al. 2007). Nonetheless, all of these chemokines are also present in human, indicating that humans have an additional point of IL-8-like chemokine function that is absent from mice. Interestingly, transgenic mice over-expressing human interleukin-8 showed an exacerbation of inflammation linked to an increased mobilisation of immature myeloid cells in DSS-induced colitis (Asfaha et al. 2013). The presence of IL-8 in human, but not mouse, might reflect the greatly increased proportion of circulating neutrophils in human blood (Doeing et al. 2003), and indicates that mouse models do not necessarily provide a full picture of neutrophil recruitment and function. IL-8 is not the only chemokine without a strict one-to-one orthologue in mouse. This is also the case for other chemokines from the CXC chemokine family, for example CXCL11, as well as CCL13, CCL14, CCL15 and CCL18 from the CC chemokine family. Conversely, some chemokines (CCL6, CCL9, CCL12, CXCL15) are present in the mouse genome, but not in human (Mestas and Hughes 2004). Furthermore, the chemokine-like receptor GPR33 is present in some rodents including mice, but is a pseudogene in human. Murine GPR33 expression is regulated by agonists of several TLRs, suggesting that it has some function in innate immunity (Bohnekamp et al. 2010). Thus, it appears to have become dispensable for immune function during evolution of humans and some other species.

The interleukin (IL) family of cytokines, a nomenclature first introduced in 1979 (Attendees of the Second International Lymphokine Workshop 1979), is generally under strong evolutionary constraint. One striking example is illustrated by the marine protozoa *Euplotes raikovi* that produces pheromones to find permissive partners for reproduction. One of these pheromones, Er-1, shares properties with human IL-2. Remarkably, Er-1 can bind to IL-2R and conversely, human IL-2 is recognised by the Er-1-binding pheromone receptor (Vallesi et al. 1998). Gene duplication within the IL family has resulted in four major groups that can be distinguished by their structural features; the IL-1-like cytokines, the class I helical cytokines (IL-4-like, γ-chain and IL6/12-like), the class II helical cytokines (IL-10-like and IL-28-like) and the IL-17-like cytokines (Brocker et al. 2010). These genes are present in distinct chromosomal clusters in both human and mouse. For example, one human cluster containing IL-3, IL-4, IL-5 and *CSF2* (Chr 5q23-32) corresponds to a syntenic region in the mouse genome at Chr 11A1 (Lee et al. 1989). Despite the evolutionary constraint in cytokine evolution, as exemplified by IL-2 and Er-1 above, there are some specific cytokines that show differences between human and mouse. For example, IL-26, an IL-10-like cytokine, is present in human but does not have an orthologue in mice or other rodents (Donnelly et al. 2010), which indicates that mouse studies will be unable to capture the function of this particular cytokine in homeostasis and

disease. In this regard, IL-26 expression was elevated in Crohn's disease patients (Dambacher et al. 2009) and has also been linked to pathological Th17 responses in rheumatoid arthritis (Corvaisier et al. 2012).

The interferons (IFNs) represent another subfamily of helicoidal interleukins, discovered in 1957 by Isaacs and Lindenmann. Interferon was first described as a soluble factor allowing cells to resist influenza infection, and therefore was named because it was 'interfering' with the spreading of the virus (Isaacs and Lindenmann 1987). IFNs were originally grouped into type I IFNs (encompassing IFN-α and -β IFNs) and type II IFNs (the IFN-γ family). Type I IFNs include 17 subtypes binding to one heterodimeric receptor (IFNAR1 and 2) (Uze et al. 2007). In 2003, a new class of IFNs (type III or IFN-λ) were discovered; these include IL-28A, IL-28B and IL-29 (Kotenko et al. 2003; Sheppard et al. 2003). IL-29 is a pseudogene in mice, suggesting that there may be functional differences between human and mouse with respect to type III IFN action. Indeed, recent work has demonstrated that human hepatocytes responded to type III IFN, whereas mouse cholangiocytes but not hepatocytes were responsive (Hermant et al. 2014). The type I IFN family also displays species differences. Human and mouse both possess α, β, ε and κ IFN, but only mice have IFN-ζ, also named limitin. This IFN, which was discovered in 2000, strongly inhibits the growth of B-lymphocyte precursors (Oritani et al. 2000). Alternatively, two other IFNs are specific to the human lineage, ω and δ. IFN ω is thought to have diverged from IFN-α around 130 MYA (Flores et al. 1991). One functional gene has been found in human (along with two pseudogenes), but only one pseudogene has been reported in mouse (Woelk et al. 2007). Finally, IFN-τ was found in ruminants during pregnancy, was reported as a pseudogene in human, but is completely absent from the mouse genome (Leaman and Roberts 1992). This IFN arose 36 MYA in ruminants and is absent from all other species, suggesting that the IFN gene family is under strong environmental selection pressure and that individual family members are frequently duplicated and develop new functions (Roberts et al. 2003; Walker and Roberts 2009).

6.4 Human Versus Mouse Differences in LPS Responses: A Case Study

Some of our specific interests in human versus mouse differences in innate immunity have focused on LPS-initiated TLR4 signalling. It has long been appreciated that there is a striking difference between human and mouse in susceptibility to LPS, with mice being substantially more resistant. Indeed it is estimated that the dose of LPS required to caused lethality in mice is approximately 10,000 more than that required to cause severe shock in human (Warren et al. 2010a). Despite this, mice are widely employed to study the pathophysiology of LPS-mediated inflammation, as well as the molecular mechanisms responsible. This prompted Robert S. Munford, a prominent LPS researcher, to describe this

difference as 'another dirty little secret' (Munford 2010) (in reference to the original use of this phrase to describe the requirements of adjuvants to activate innate immunity, as a component of effective vaccines). Intriguingly, while TLR4 is well-recognised as the cell-surface receptor for LPS (Akashi et al. 2003), it is now known that a cytoplasmic LPS receptor also initiates caspase-11 activation in mouse macrophages (Hagar et al. 2013). As described above, caspase-11 has no strict one-to-one orthologue in human, instead being similar to both caspase-4 and caspase-5. Thus, this immediately raises the possibility of signalling differences between human and mouse upon sensing cytoplasmic LPS. Below, we focus on human versus mouse differences in sensing extracellular LPS via TLR4. Although we describe some differences in proximal signalling events, we primarily focus on differences in regulated gene expression and the mechanisms for this.

6.4.1 The Role of TLR4 in Innate Immunity

Because of a spontaneous genetic mutation, the C3H/HeJ mouse strain is insensitive to endotoxin (Verghese et al. 1980) and the gene responsible was eventually mapped to TLR4 by Beutler and colleagues (Poltorak et al. 1998). Soon after, Hoshino et al. (1999) confirmed the essential role of this receptor in LPS detection using TLR4 knockout mice. However, TLR4 alone is not sufficient to transduce cell signalling in response to LPS; rather, it works in concert with CD14 and MD-2. The CD14 co-receptor (present as either a GPI-anchored membrane protein or as a soluble protein) sensitises responses to LPS (Pugin et al. 1993; Viriyakosol et al. 2000). LPS, captured by LPS-binding protein, binds to CD14 which in turn transfers it to the TLR4–MD-2 complex. MD-2 is important for LPS/TLR4 binding as it has a β-cup fold, usually involved in lipid storage, and this is important for interaction with the acyl-chain of lipid A (Park and Lee 2013) (see below). The importance of CD14 has been confirmed using CD14-deficient mice which are highly resistant to LPS challenge (Haziot et al. 1996). Similarly, MD-2 deficiency leads to hyporesponsiveness to LPS (Visintin et al. 2006). Activation of TLR4 triggers an intracellular signalling cascade involving the recruitment of adaptor molecules (e.g. MYD88), followed by activation of kinases (e.g. IRAK-1), ubiquitin ligases (e.g. TRAF6) and transcription factors (e.g. NF-κB) (Kawai and Akira 2007). Interestingly, while genetic and structural evidence irrefutably identifies TLR4 as an LPS receptor, this PRR has also been implicated in the detection of an ever-increasing list of other pathogen and host molecules, many of which show little structural relationship with lipid A (the component of LPS recognised by TLR4). Such differences, along with the fact that LPS is a common contaminant in biological preparations, necessitate extreme caution when interpreting such observations. Nonetheless, TLR4 has been reported to detect various other microbial products (e.g. the fusion protein of respiratory syncytial virus, proteins from the viral envelope of mouse mammary tumour virus (MMTV) and moloney murine leukaemia virus (MMLV), glycoinositol phospholipids of *Trypanosoma cruzi*, and the polysaccharide of *Cryptococcus neoformans* and *Candida albicans* (Kurt-Jones et al.

2000; Shoham et al. 2001; Tada et al. 2002; Oliveira et al. 2004)), as well as host-derived danger signals (e.g. HMGB1, the surfactant protein SP-A and Fetuin-A (Yu et al. 2006; Guillot et al. 2002; Pal et al. 2012)).

6.4.2 Human Versus Mouse Differences in Ligand Recognition by TLR4

The critical motifs within LPS that are required for detection by TLR4 are six acyl chains and two phosphate groups that are both present on the disaccharide of the lipid A component. This detection mechanism has been exploited by some Gram-negative pathogens, which can alter these moieties to avoid detection by TLR4 (Caroff et al. 2002). Some of these effects impact on recognition by human TLR4 versus mouse TLR4. For example, at higher temperatures, the LPS of *Yersinia pestis* becomes hypoacylated (Knirel et al. 2005). This LPS is less stimulatory to human compared to murine cells, whereas LPS from *Yersinia pestis* (*Y. pestis*) that has the normal level of acylation is similarly detected by human and mouse cells. This difference can be attributed to recognition by TLR4/MD-2, as evidenced by studies with TLR4-/MD2-humanised mice (Hajjar et al. 2012). The capacity of mouse to sense all forms of *Y. pestis* LPS likely contributes to its relative resistance to this pathogen (Lambert et al. 2011). Similar differences have been reported between human and mouse TLR4 in their capacity to sense penta-acylated LPS from *Neisseria meningitidis* (Steeghs et al. 2008). Human and mouse TLR4 also show differential capacity to recognise non-LPS ligands and initiate downstream signalling. For example, human TLR4 mediates hypersensitivity to nickel, whereas mouse TLR4 does not. These differences were linked to two histidines (456 and 458) of human TLR4 that are not conserved in murine TLR4 (Schmidt et al. 2010). Taxol, an anticancer agent that is also an LPS mimetic (Byrd-Leifer et al. 2001), is recognised by the mouse TLR4-MD-2 complex, but not the human one. This selective ligand-specificity was linked to the presence of a glutamine residue in the mouse sequence, which is absent in human TLR4 (Kawasaki et al. 2001). These represent just some of the many differences in TLR4/MD-2 between human and mouse that impact on ligand recognition.

6.4.3 Human Versus Mouse Differences in LPS-Regulated Gene Expression

Although there is strong conservation in TLR4 intracellular signalling pathways in human and mouse, several studies have identified substantial differences in downstream TLR4 target genes. Indeed, dramatic divergence in human versus mouse gene expression profiles during inflammatory shock (trauma, endotoxaemia, burn) has been reported (Seok et al. 2013). This study reported that the

correlation between human and mouse gene expression programmes in these states was close to random (R^2 between 0.0 and 0.1), though it must be considered that a range of other factors makes such in vivo comparisons very difficult. Additionally, we have also reported substantial divergence, at a cellular level, in human and mouse transcriptional responses to LPS (Schroder et al. 2012). We observed striking differences in the LPS-regulated expression of numerous inputs (e.g. cytokine receptors such as IL-7R), outputs (e.g. secreted molecules such as the chemokine CCL20), feedback regulators that control signalling (e.g. *IRAK3*, *JDP2*, *SOCS1*, *ATF3*) and direct antimicrobial effector genes (e.g. iNOS, IDO). The differential regulation of the latter is of particular interest, given that some of these effects likely reflect host–pathogen co-evolution and that some of these differences are likely to be particularly important in the context of host defence.

6.4.3.1 Human Versus Mouse Differences in TLR4-Regulated Antimicrobial Effector Genes

While innate immune cells are armed with a suite of antimicrobial weapons that are immediately engaged upon phagocytosis of microorganisms (e.g. degradative enzymes, reactive oxygen species), others are induced in a delayed fashion, often downstream of TLRs. These are likely important for defence against pathogens that evade primary antimicrobial responses. A classic example is provided by inducible nitric oxide synthases (iNOS or NOS2), a member of the NOS family of enzymes. iNOS is robustly induced by LPS in mouse macrophages via autocrine type I IFN signalling (Fujihara et al. 1994), and oxidises L-arginine to L-citrulline, with nitric oxide (NO) generated as a co-product. Formation of reactive nitrogen species through iNOS serves as a defence mechanism by inflicting oxidative damage on intraphagosomal microbes resulting in inactivation of proteins and DNA damage. Nitric oxide can also contribute to host defence by directly regulating bacterial gene expression. During *Salmonella* infection in mouse macrophages, NO production down-regulates the Salmonella pathogenicity island-2 (SPI2), which normally allows the bacteria to evade phagolysosomal degradation (McCollister et al. 2005). Consequently, mice genetically deficient in iNOS show increased susceptibility to various intramacrophage pathogens including *Leishmania major*, *Mycobacterium* species and *Salmonella enterica* (Wei et al. 1995; MacMicking et al. 1997; Alam et al. 2002). Despite these mouse studies, the importance of iNOS for host defence in humans is still unclear (Weinberg 1998). This gene is not TLR4-inducible in human macrophages (Schroder et al. 2012), as it is in the mouse. This does not mean that it has no function in host defence in human, rather the cell types expressing it and/or the stimuli that induce it may differ. Thus, it may be that iNOS functions as part of innate immunity in both human and mouse, but that it functions in defence against different pathogens in these two species. Indeed, studies on the human iNOS promoter do predict a role in control of infectious diseases. One study associated iNOS promoter polymorphisms with malaria outcome; children with the -954 G > C SNP and the

microsatellite repeats CCTTT(x8) were more protected against hyperparasitaemia, whereas those with −1,173 C > T SNP and CCTTT(x13) showed higher fatality rates (Cramer et al. 2004). Another study also reported associations between iNOS promoter variants and tuberculosis susceptibility in African-Americans (Velez et al. 2009).

In contrast to the mouse-specific induction of iNOS by LPS, the gene encoding the enzyme indoleamine 2,3-dioxygenase (IDO) is inducible by LPS in human, but not mouse (Roshick et al. 2006). This enzyme metabolises L-tryptophan, thus depriving certain intracellular pathogens of this essential amino acid (Zelante et al. 2009). Unlike mouse macrophages, IFN-γ pretreated human macrophages induce IDO during infection resulting in tryptophan depletion (Murray et al. 1989). This divergence between human and mouse may be interconnected with differential iNOS regulation, since NO can inhibit IDO expression (Thomas et al. 1994). Thus, TLR4-inducible iNOS expression, as occurs in mouse macrophages, would be predicted to prevent inducible IDO expression. This differential use of IDO and tryptophan starvation can result in interesting effects on host–pathogen dynamics. For example, although different species of *Chlamydia* can infect both mouse and human cells, only the human-tropic species express their own tryptophan synthetase gene in order to circumvent the host tryptophan starvation pathway (Nelson et al. 2005). Several other divergently regulated TLR4 target genes that we have identified are also likely to have roles in direct antimicrobial responses (Schroder et al. 2012), suggesting that many other differences between human and mouse exist in this respect.

6.4.3.2 Mechanisms of Divergence in TLR4-Regulated Gene Expression

Some of the divergently regulated TLR4 target genes that are well-validated in the literature, such as iNOS, have been studied to identify the molecular basis for their divergent regulation. For example, promoter-reporter analysis demonstrated that the human iNOS promoter did not confer LPS responsiveness. Zhang et al. (1996) found that mouse promoter elements required for induction were not conserved within the human promoter. They reported the presence of multiple substitutions in an enhancer element of the human promoter, in particular in the ISRE and GAS elements within region II. Subsequent gene knock-out studies demonstrated essential roles for the transcription factors STAT1 (Ohmori and Hamilton 2001) and IRF-1 (Kamijo et al. 1994) for inducible expression of iNOS in mouse macrophages. The iNOS gene therefore provides an example of how specific promoter sequences can dramatically alter LPS responsiveness in human versus mouse. Very surprisingly, in our own studies (Schroder et al. 2012), we actually found that promoters of divergently regulated LPS target genes were much more highly conserved across species than those of non-divergently regulated LPS target genes. Furthermore, these genes were typically associated with a specific promoter architecture (enriched for TATA boxes and depleted of CpG islands) and were the

most dramatically regulated of the LPS target genes. This promoter architecture is consistent with the genes being subjected to high levels of regulatory inputs (i.e. multiple transcription factors acting through multiple binding sites). From these findings, we inferred that these particular TLR4 target genes must be functionally important and highly constrained as evidenced by promoter conservation across multiple mammalian species, but at the same time very susceptible to evolutionary change in expression due to the nature of their promoter architecture. That is, gene regulation that depends on multiple regulatory inputs would mean that a change in the binding capacity of one transcription factor (e.g. through evolutionary loss of a binding site or through loss of expression of a particular transcription factor) would result in a dramatic change in gene expression. With regard to loss of binding sites, examples of this are apparent with the iNOS gene (above) and the IL-7R gene (below). With respect to differences in transcription factor expression that might contribute to such effects, we would predict that differences in basal transcription factor expression would be particularly relevant. Nonetheless, we also identified a total of 49 transcription factors that were themselves differentially LPS regulated between the species. For example, the transcription factors HEY1 and HESX1 were robustly induced by LPS in human, but not mouse, macrophages. Thus, our current model is that, seemingly paradoxically, divergently regulated promoters show a high level of evolutionary constraint, but the complex nature of their promoter architecture means that small evolutionary changes in promoter sequence or available transcription factors can have dramatic effects on regulated gene expression.

6.4.3.3 IL-7R: An Example of a 'Human-Specific' TLR4 Target Gene

One recently described example of a human-specific TLR4 target gene is the IL-7R; we showed that this gene, as well as surface IL-7R protein, was upregulated by LPS in macrophages from human, but not mouse (Schroder et al. 2012). IL-7R is a heterodimer of one specific IL-7Rα chain (CD127) and the γ-chain, which is common to IL-2, 4, 9, 12 and 21. Its activation induces phosphorylation of tyrosine residues in the cytoplasmic domain and leads to intracellular signalling through Jak1, STAT5, PI3K and src kinases (Hofmeister et al. 1999). Its ligand, IL-7, was discovered in 1988 and was first named lymphopoietin-1 (LP-1), due to its action in promoting B cell development (Namen et al. 1988b). This cytokine was cloned shortly after by Raymond Goodwin's team from a murine stromal cell line derived from bone marrow (Namen et al. 1988a). IL-7 is an anti-apoptotic cytokine, which has a key role in survival of memory lymphocytes and adaptive immune functions. For these reasons, IL-7 is currently in clinical trials as a treatment for cancer and viral infections (Hotchkiss and Opal 2010). IL-7R deficiency causes severe immune deficiency in humans (Mazzucchelli et al. 2012) and aberrant IL-7R activation is associated with chronic inflammatory disease, for example rheumatoid arthritis (Pickens et al. 2011). Indeed, an anti-IL-7 antibody was therapeutic in a mouse collagen-induced arthritis model (Chen et al. 2013). Pathophysiological

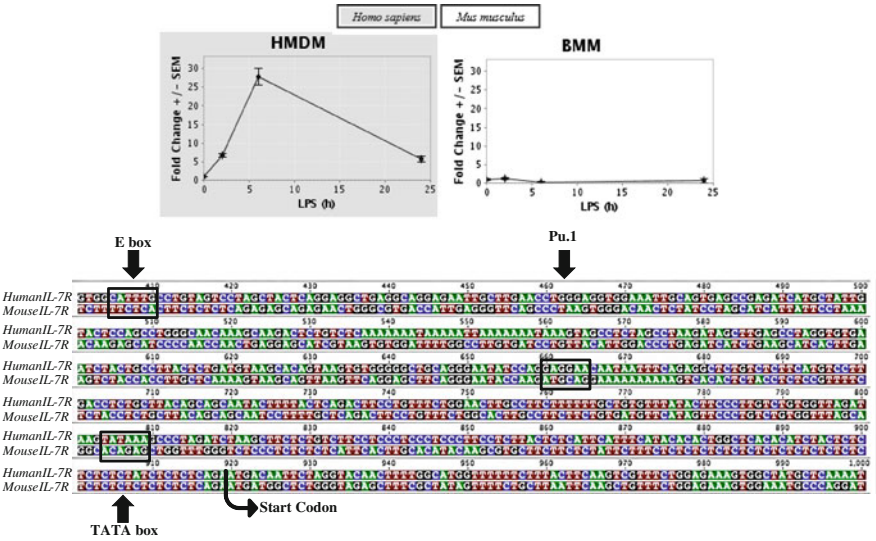


Fig. 6.2 IL-7R *is a 'human-specific' TLR4 target gene*. IL-7R is up-regulated in primary human macrophages (human monocyte-derived macrophages, HMDM), but not primary mouse macrophages (bone marrow-derived macrophages, BMM) in response to LPS. Alignment of the human and mouse IL-7R promoter regions shows differences between species. In particular, a candidate PU.1 binding site and a TATA box are present in the human promoter, but are absent from that of the mouse. In addition, 1 possible enhancer consensus sequence (E-box) is present only in the human promoter. Gene expression data of IL-7R were extracted from Macgate (http://www.macgate.qfab.org). Human and mouse sequences were aligned using the software Geneious

effects of IL-7 may not only relate to effects on lymphocytes, but also to effects on endothelial cells and innate immune cells such as monocytes and macrophages. For example, IL-7 up-regulated cell adhesion molecules on endothelial cells and induced the recruitment of monocytes/macrophages (Li et al. 2012). Interestingly, mouse IL-7 is active on human IL-7R^{+} cells indicating that there is species cross-reactivity at the level of the receptor (Barata et al. 2006).

Regarding our own work, we found that LPS up-regulated expression of IL-7R in primary human macrophages, whereas this response was not observed in primary mouse macrophages (Fig. 6.2). In the case of this gene, genetic studies would suggest that variation within the promoter region can have a profound impact on gene expression. For example, Teutsch et al. (2003) discovered 11 polymorphisms of IL7R associated with multiple sclerosis, of which three are localised within the promoter region. The homology between the human and the murine IL-7Rα coding regions (cDNA) is close to 64 % (www.ensembl.org). In relation to specific regulatory factors, De Koter et al. (2007) showed that the transcription factors PU.1 and GA-binding protein (GABP) can activate the human IL-7R promoter through a highly conserved Ets binding site. Interestingly, in our comparisons of the human and mouse IL-7R promoters, we observed that a binding site for the transcription factor PU.1, a key regulator of myeloid

development, was present in the human promoter, but not that of the mouse (Fig. 6.2). Even more strikingly, a TATA box (TATAA) that was present in the human proximal promoter was also absent from mouse. Given that TATA box-containing promoters are typically dynamically regulated, this difference is likely to account for the failure of LPS to up-regulate the IL-7R gene in mouse macrophages.

6.5 Conclusions

In summary, innate immunity is an inherently plastic system that is exquisitely sensitive to evolutionary change. Studying species differences in innate immunity thus provides an opportunity to understand mechanisms driving evolutionary divergence, the roles of species-specific innate immune responses in host defence, and possible limitations of specific model organisms for understanding human disease processes. For these reasons, studies on human versus mouse differences in innate immunity will continue to provide important insights and perspectives.

Acknowledgments RK is supported by an Australian Research Council DECRA Fellowship (DE1310470), and MJS is supported by an Australian Research Council Future Fellowship (FT100100657) and an honorary NHMRC Senior Research Fellowship (APP1003470). This work was also supported by a grant from the National Health and Medical Research Council of Australia (ID631531).

References

Akashi S, Saitoh S, Wakabayashi Y, Kikuchi T, Takamura N, Nagai Y, Kusumoto Y, Fukase K, Kusumoto S, Adachi Y, Kosugi A, Miyake K (2003) Lipopolysaccharide interaction with cell surface Toll-like receptor 4-MD-2: higher affinity than that with MD-2 or CD14. J Exp Med 198(7):1035–1042. doi:10.1084/jem.20031076

Alam MS, Akaike T, Okamoto S, Kubota T, Yoshitake J, Sawa T, Miyamoto Y, Tamura F, Maeda H (2002) Role of nitric oxide in host defense in murine salmonellosis as a function of its antibacterial and antiapoptotic activities. Infect Immun 70(6):3130–3142

Alexopoulou L, Holt AC, Medzhitov R, Flavell RA (2001) Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. Nature 413(6857):732–738

Ambrose CT (2006) The Osler slide, a demonstration of phagocytosis from 1876 Reports of phagocytosis before Metchnikoff's 1880 paper. Cell Immunol 240(1):1–4

Antonopoulos C, El Sanadi C, Kaiser WJ, Mocarski ES, Dubyak GR (2013) Proapoptotic chemotherapeutic drugs induce noncanonical processing and release of IL-1beta via caspase-8 in dendritic cells. J Immunol 191(9):4789–4803. doi:10.4049/jimmunol.1300645

Ariffin JK, Sweet MJ (2013) Differences in the repertoire, regulation and function of Toll-like receptors and inflammasome-forming nod-like receptors between human and mouse. Curr Opin Microbiol 16(3):303–310. doi:10.1016/j.mib.2013.03.002

Asfaha S, Dubeykovskiy AN, Tomita H, Yang X, Stokes S, Shibata W, Friedman RA, Ariyama H, Dubeykovskaya ZA, Muthupalani S, Ericksen R, Frucht H, Fox JG, Wang TC (2013) Mice that express human interleukin-8 have increased mobilization of immature myeloid cells, which exacerbates inflammation and accelerates colon carcinogenesis. Gastroenterology 144(1):155–166. doi:10.1053/j.gastro.2012.09.057

Atianand MK, Duffy EB, Shah A, Kar S, Malik M, Harton JA (2011) Francisella tularensis reveals a disparity between human and mouse NLRP3 inflammasome activation. J Biol Chem 286(45):39033–39042. doi:10.1074/jbc.M111.244079

Attendees of the Second International Lymphokine Workshop (1979) Revised nomenclature for antigen-nonspecific T cell proliferation and helper factors. J Immunol 123 (6):2928-2929

Bals R, Weiner DJ, Moscioni AD, Meegalla RL, Wilson JM (1999) Augmentation of innate host defense by expression of a cathelicidin antimicrobial peptide. Infect Immun 67(11):6084–6089

Barata JT, Silva A, Abecasis M, Carlesso N, Cumano A, Cardoso AA (2006) Molecular and functional evidence for activity of murine IL-7 on human lymphocytes. Exp Hematol 34(9):1133–1142. doi:10.1016/j.exphem.2006.05.001

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, Kidd JR, Kidd KK, Alcais A, Ragimbeau J, Pellegrini S, Abel L, Casanova JL, Quintana-Murci L (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet 5(7):e1000562. doi:10.1371/journal.pgen.1000562

Bauer S, Kirschning CJ, Hacker H, Redecke V, Hausmann S, Akira S, Wagner H, Lipford GB (2001) Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. Proc Natl Acad Sci USA 98(16):9237–9242

Bekpen C, Hunn JP, Rohde C, Parvanova I, Guethlein L, Dunn DM, Glowalla E, Leptin M, Howard JC (2005) The interferon-inducible p47 (IRG) GTPases in vertebrates: loss of the cell autonomous resistance mechanism in the human lineage. Genome Biol 6(11):R92. doi:10.1186/gb-2005-6-11-r92

Bohnekamp J, Boselt I, Saalbach A, Tonjes A, Kovacs P, Biebermann H, Manvelyan HM, Polte T, Gasperikova D, Lkhagvasuren S, Baier L, Stumvoll M, Rompler H, Schoneberg T (2010) Involvement of the chemokine-like receptor GPR33 in innate immunity. Biochem Biophys Res Commun 396(2):272–277. doi:10.1016/j.bbrc.2010.04.077

Boldajipour B, Doitsidou M, Tarbashevich K, Laguri C, Yu SR, Ries J, Dumstrei K, Thelen S, Dorries J, Messerschmidt EM, Thelen M, Schwille P, Brand M, Lortat-Jacob H, Raz E (2011) Cxcl12 evolution–subfunctionalization of a ligand through altered interaction with the chemokine receptor. Development 138(14):2909–2914. doi:10.1242/dev.068379

Bordon Y (2012) Innate immunity: TLR13, unlucky, but just for some. Nat Rev Immunol 12(9):618–619. doi:10.1038/nri3284

Brinkman-Van der Linden EC, Sjoberg ER, Juneja LR, Crocker PR, Varki N, Varki A (2000) Loss of N-glycolylneuraminic acid in human evolution. Implications for sialic acid recognition by siglecs. J Biol Chem 275(12):8633–8640

Brocker C, Thompson D, Matsumoto A, Nebert DW, Vasiliou V (2010) Evolutionary divergence and functions of the human interleukin (IL) gene family. Human Genomics 5(1):30–55

Brunette RL, Young JM, Whitley DG, Brodsky IE, Malik HS, Stetson DB (2012) Extensive evolutionary and functional diversity among mammalian AIM2-like receptors. J Exp Med 209(11):1969–1983. doi:10.1084/jem.20121960

Byrd-Leifer CA, Block EF, Takeda K, Akira S, Ding A (2001) The role of MyD88 and TLR4 in the LPS-mimetic activity of taxol. Eur J Immunol 31(8):2448–2457. doi:10.1002/1521-4141(200108)31:8<2448:AID-IMMU2448>3.0.CO;2-N

Caroff M, Karibian D, Cavaillon JM, Haeffner-Cavaillon N (2002) Structural and functional analyses of bacterial lipopolysaccharides. Microbes Infect 4(9):915–926

Chamaillard M, Hashimoto M, Horie Y, Masumoto J, Qiu S, Saab L, Ogura Y, Kawasaki A, Fukase K, Kusumoto S, Valvano MA, Foster SJ, Mak TW, Nunez G, Inohara N (2003) An essential role for NOD1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. Nat Immunol 4(7):702–707

Chen Z, Kim SJ, Chamberlain ND, Pickens SR, Volin MV, Volkov S, Arami S, Christman JW, Prabhakar BS, Swedler W, Mehta A, Sweiss N, Shahrara S (2013) The novel role of IL-7 ligation to IL-7 receptor in myeloid cells of rheumatoid arthritis and collagen-induced arthritis. J Immunol 190(10):5256–5266. doi:10.4049/jimmunol.1201675

Chou HH, Takematsu H, Diaz S, Iber J, Nickerson E, Wright KL, Muchmore EA, Nelson DL, Warren ST, Varki A (1998) A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc Natl Acad Sci USA 95(20):11751–11756

Chu H, Pazgier M, Jung G, Nuccio SP, Castillo PA, de Jong MF, Winter MG, Winter SE, Wehkamp J, Shen B, Salzman NH, Underwood MA, Tsolis RM, Young GM, Lu W, Lehrer RI, Baumler AJ, Bevins CL (2012) Human alpha-defensin 6 promotes mucosal innate immunity through self-assembled peptide nanonets. Science 337(6093):477–481. doi:10.1126/science.1218831

Chun HJ, Zheng L, Ahmad M, Wang J, Speirs CK, Siegel RM, Dale JK, Puck J, Davis J, Hall CG, Skoda-Smith S, Atkinson TP, Straus SE, Lenardo MJ (2002) Pleiotropic defects in lymphocyte activation caused by caspase-8 mutations lead to human immunodeficiency. Nature 419(6905):395–399. doi:10.1038/nature01063

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, Mouse Genome Sequencing C (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7(5):e1000112. doi:10.1371/journal.pbio.1000112

Colmenares M, Puig-Kroger A, Pello OM, Corbi AL, Rivas L (2002) Dendritic cell (DC)-specific intercellular adhesion molecule 3 (ICAM-3)-grabbing nonintegrin (DC-SIGN, CD209), a C-type surface lectin in human DCs, is a receptor for Leishmania amastigotes. J Biol Chem 277(39):36766–36769. doi:10.1074/jbc.M205270200

Cortopassi GA, Wilson AC (1990) Recent origin of the P lysozyme gene in mice. Nucleic Acids Res 18 (7):1911

Corvaisier M, Delneste Y, Jeanvoine H, Preisser L, Blanchard S, Garo E, Hoppe E, Barre B, Audran M, Bouvard B, Saint-Andre JP, Jeannin P (2012) IL-26 is overexpressed in rheumatoid arthritis and induces proinflammatory cytokine production and Th17 cell generation. PLoS Biol 10(9):e1001395. doi:10.1371/journal.pbio.1001395

Cramer JP, Mockenhaupt FP, Ehrhardt S, Burkhardt J, Otchwemah RN, Dietz E, Gellert S, Bienzle U (2004) iNOS promoter variants and severe malaria in Ghanaian children. Tropical medicine and international health: TM and IH 9(10):1074–1080. doi:10.1111/j.1365-3156.2004.01312.x

Cridland JA, Curley EZ, Wykes MN, Schroder K, Sweet MJ, Roberts TL, Ragan MA, Kassahn KS, Stacey KJ (2012) The mammalian PYHIN gene family: phylogeny, evolution and expression. BMC Evol Biol 12:140. doi:10.1186/1471-2148-12-140

Crocker PR, Jefferies WA, Clark SJ, Chung LP, Gordon S (1987) Species heterogeneity in macrophage expression of the CD4 antigen. J Exp Med 166(2):613–618

Cunliffe RN, Rose FR, Keyte J, Abberley L, Chan WC, Mahida YR (2001) Human defensin 5 is stored in precursor form in normal Paneth cells and is expressed by some villous epithelial cells and by metaplastic Paneth cells in the colon in inflammatory bowel disease. Gut 48(2):176–185

Dambacher J, Beigel F, Zitzmann K, De Toni EN, Goke B, Diepolder HM, Auernhammer CJ, Brand S (2009) The role of the novel Th17 cytokine IL-26 in intestinal inflammation. Gut 58(9):1207–1217. doi:10.1136/gut.2007.130112

Dawson HD, Loveland JE, Pascal G, Gilbert JG, Uenishi H, Mann KM, Sang Y, Zhang J, Carvalho-Silva D, Hunt T, Hardy M, Hu Z, Zhao SH, Anselmo A, Shinkai H, Chen C, Badaoui B, Berman D, Amid C, Kay M, Lloyd D, Snow C, Morozumi T, Cheng RP, Bystrom M, Kapetanovic R, Schwartz JC, Kataria R, Astley M, Fritz E, Steward C, Thomas M, Wilming L, Toki D, Archibald AL, Bed'hom B, Beraldi D, Huang TH, Ait-Ali T, Blecha F, Botti S, Freeman TC, Giuffra E, Hume DA, Lunney JK, Murtaugh MP, Reecy JM, Harrow JL, Rogel-Gaillard C, Tuggle CK (2013) Structural and functional annotation of the porcine immunome. BMC Genom 14:332. doi:10.1186/1471-2164-14-332

DeKoter RP, Schweitzer BL, Kamath MB, Jones D, Tagoh H, Bonifer C, Hildeman DA, Huang KJ (2007) Regulation of the interleukin-7 receptor alpha promoter by the Ets transcription factors

PU.1 and GA-binding protein in developing B cells. J Biol Chem 282(19):14194–14204. doi:10.1074/jbc.M700377200

Demetrius L (2005) Of mice and men. When it comes to studying ageing and the means to slow it down, mice are not just small humans. EMBO reports 6 Spec No:S39–S44. doi:10.1038/sj.embor.7400422

Desnues B, Macedo AB, Roussel-Queval A, Bonnardel J, Henri S, Demaria O, Alexopoulou L (2014) TLR8 on dendritic cells and TLR9 on B cells restrain TLR7-mediated spontaneous autoimmunity in C57BL/6 mice. Proc Natl Acad Sci USA 111(4):1497–1502. doi:10.1073/pnas.1314121111

DeYoung BJ, Innes RW (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol 7(12):1243–1249

Doeing DC, Borowicz JL, Crockett ET (2003) Gender dimorphism in differential peripheral blood leukocyte counts in mice using cardiac, tail, foot, and saphenous vein puncture methods. BMC clinical pathology 3(1):3. doi:10.1186/1472-6890-3-3

Donnelly RP, Sheikh F, Dickensheets H, Savan R, Young HA, Walter MR (2010) Interleukin-26: an IL-10-related cytokine produced by Th17 cells. Cytokine Growth Factor Rev 21(5):393–401. doi:10.1016/j.cytogfr.2010.09.001

Dorus S, Gilbert SL, Forster ML, Barndt RJ, Lahn BT (2003) The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. Hum Mol Genet 12(14):1643–1650

Eisenacher K, Krug A (2012) Regulation of RLR-mediated innate immune signaling–it is all about keeping the balance. Eur J Cell Biol 91(1):36–47. doi:10.1016/j.ejcb.2011.01.011

Eisenbarth SC, Williams A, Colegio OR, Meng H, Strowig T, Rongvaux A, Henao-Mejia J, Thaiss CA, Joly S, Gonzalez DG, Xu L, Zenewicz LA, Haberman AM, Elinav E, Kleinstein SH, Sutterwala FS, Flavell RA (2012) NLRP10 is a NOD-like receptor essential to initiate adaptive immunity by dendritic cells. Nature 484(7395):510–513. doi:10.1038/nature11012

Eisenhauer PB, Lehrer RI (1992) Mouse neutrophils lack defensins. Infect Immun 60(8):3446–3447

Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, Knight JC (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 343(6175):1246949. doi:10.1126/science.1246949

Fan X, Patera AC, Pong-Kennedy A, Deno G, Gonsiorek W, Manfra DJ, Vassileva G, Zeng M, Jackson C, Sullivan L, Sharif-Rodriguez W, Opdenakker G, Van Damme J, Hedrick JA, Lundell D, Lira SA, Hipkin RW (2007) Murine CXCR1 is a functional receptor for GCP-2/CXCL6 and interleukin-8/CXCL8. J Biol Chem 282(16):11658–11666. doi:10.1074/jbc.M607705200

Fernandes-Alnemri T, Yu JW, Juliana C, Solorzano L, Kang S, Wu J, Datta P, McCormick M, Huang L, McDermott E, Eisenlohr L, Landel CP, Alnemri ES (2010) The AIM2 inflammasome is critical for innate immunity to Francisella tularensis. Nat Immunol 11(5):385–393. doi:10.1038/ni.1859

Fleming A (1922) On a remarkable bacteriolytic element found in tissues and secretions. Proc R Soc Lond B Biol Sci 93:306–317

Flores I, Mariano TM, Pestka S (1991) Human interferon omega (omega) binds to the alpha/beta receptor. J Biol Chem 266(30):19875–19877

Fujihara M, Ito N, Pace JL, Watanabe Y, Russell SW, Suzuki T (1994) Role of endogenous interferon-beta in lipopolysaccharide-triggered activation of the inducible nitric-oxide synthase gene in a mouse macrophage cell line, J774. J Biol Chem 269(17):12773–12778

Gayon J, Burian RM (2000) France in the era of Mendelism (1900–1930). C R Acad Sci 323(12):1097–1106

Gazzinelli RT, Mendonca-Neto R, Lilue J, Howard J, Sher A (2014) Innate resistance against toxoplasma gondii: an Evolutionary tale of mice, cats, and men. Cell Host Microbe 15(2):132–138. doi:10.1016/j.chom.2014.01.004

Ghosh D, Porter E, Shen B, Lee SK, Wilk D, Drazba J, Yadav SP, Crabb JW, Ganz T, Bevins CL (2002) Paneth cell trypsin is the processing enzyme for human defensin-5. Nat Immunol 3(6):583–590. doi:10.1038/ni797

Girardin SE, Boneca IG, Carneiro LA, Antignac A, Jehanno M, Viala J, Tedin K, Taha MK, Labigne A, Zahringer U, Coyle AJ, DiStefano PS, Bertin J, Sansonetti PJ, Philpott DJ (2003a) Nod1 detects a unique muropeptide from gram-negative bacterial peptidoglycan. Science 300(5625):1584–1587

Girardin SE, Boneca IG, Viala J, Chamaillard M, Labigne A, Thomas G, Philpott DJ, Sansonetti PJ (2003b) Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. J Biol Chem 278(11):8869–8872

Goldstein JL (2001) Laskers for 2001: knockout mice and test-tube babies. Nat Med 7(10):1079–1080. doi:10.1038/nm1001-1079

Gordon S (2008) Elie Metchnikoff: father of natural immunity. Eur J Immunol 38(12):3257–3264. doi:10.1002/eji.200838855

Gordon S, Todd J, Cohn ZA (1974) In vitro synthesis and secretion of lysozyme by mononuclear phagocytes. J Exp Med 139(5):1228–1248

Griffith JW, Sokol CL, Luster AD (2014) Chemokines and chemokine receptors: positioning cells for host defense and immunity. Annu Rev Immunol 32:659–702. doi:10.1146/annurev-immunol-032713-120145

Guan Y, Ranoa DR, Jiang S, Mutha SK, Li X, Baudry J, Tapping RI (2010) Human TLRs 10 and 1 share common mechanisms of innate immune sensing but not signaling. J Immunol 184(9):5094–5103. doi:10.4049/jimmunol.0901888

Guillot L, Balloy V, McCormack FX, Golenbock DT, Chignard M, Si-Tahar M (2002) Cutting edge: the immunostimulatory activity of the lung surfactant protein-A involves Toll-like receptor 4. J Immunol 168(12):5989–5992

Gurung P, Anand PK, Malireddi RK, Vande Walle L, Van Opdenbosch N, Dillon CP, Weinlich R, Green DR, Lamkanfi M, Kanneganti TD (2014) FADD and caspase-8 mediate priming and activation of the canonical and noncanonical Nlrp3 inflammasomes. J Immunol 192(4):1835–1846. doi:10.4049/jimmunol.1302839

Ha HJ, Kim DS, Hahn Y (2009) A 2.7-kb deletion in the human NLRP10 gene exon 2 occurred after the human-chimpanzee divergence. Biochem Genet 47(9–10):665–670. doi:10.1007/s10528-009-9262-2

Hagar JA, Miao EA (2014) Detection of cytosolic bacteria by inflammatory caspases. Curr Opin Microbiol 17C:61–66. doi:10.1016/j.mib.2013.11.008

Hagar JA, Powell DA, Aachoui Y, Ernst RK, Miao EA (2013) Cytoplasmic LPS activates caspase-11: implications in TLR4-independent endotoxic shock. Science 341(6151):1250–1253. doi:10.1126/science.1240988

Hajjar AM, Ernst RK, Fortuno ES 3rd, Brasfield AS, Yam CS, Newlon LA, Kollmann TR, Miller SI, Wilson CB (2012) Humanized TLR4/MD-2 mice reveal LPS recognition differentially impacts susceptibility to Yersinia pestis and Salmonella enterica. PLoS Pathog 8(10):e1002963. doi:10.1371/journal.ppat.1002963

Hasan U, Chaffois C, Gaillard C, Saulnier V, Merck E, Tancredi S, Guiet C, Briere F, Vlach J, Lebecque S, Trinchieri G, Bates EE (2005) Human TLR10 is a functional receptor, expressed by B cells and plasmacytoid dendritic cells, which activates gene transcription through MyD88. J Immunol 174(5):2942–2950

Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, Eng JK, Akira S, Underhill DM, Aderem A (2001) The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nature 410(6832):1099–1103

Haziot A, Ferrero E, Kontgen F, Hijiya N, Yamamoto S, Silver J, Stewart CL, Goyert SM (1996) Resistance to endotoxin shock and reduced dissemination of gram-negative bacteria in CD14-deficient mice. Immunity 4(4):407–414

Hd K (1918) Studies on inbreeding. 1. The effects of inbreeding on the growth and variability in the body weight of the albino rat. J Exp Zool 26:1–54

Heil F, Hemmi H, Hochrein H, Ampenberger F, Kirschning C, Akira S, Lipford G, Wagner H, Bauer S (2004) Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. Science 303(5663):1526–1529

Heinz S, Haehnel V, Karaghiosoff M, Schwarzfischer L, Muller M, Krause SW, Rehli M (2003) Species-specific regulation of Toll-like receptor 3 genes in men and mice. J Biol Chem 278(24):21502–21509. doi:10.1074/jbc.M301476200

Henriksson JT, McDermott AM, Bergmanson JP (2009) Dimensions and morphology of the cornea in three strains of mice. Invest Ophthalmol Vis Sci 50(8):3648–3654. doi:10.1167/iovs.08-2941

Hermant P, Demarez C, Mahlakoiv T, Staeheli P, Meuleman P, Michiels T (2014) Human but not mouse hepatocytes respond to interferon-lambda in vivo. PLoS ONE 9(1):e87906. doi:10.1371/journal.pone.0087906

Hofmeister R, Khaled AR, Benbernou N, Rajnavolgyi E, Muegge K, Durum SK (1999) Interleukin-7: physiological roles and mechanisms of action. Cytokine Growth Factor Rev 10(1):41–60

Hoshino K, Takeuchi O, Kawai T, Sanjo H, Ogawa T, Takeda Y, Takeda K, Akira S (1999) Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. J Immunol 162(7):3749–3752

Hotchkiss RS, Opal S (2010) Immunotherapy for sepsis–a new approach against an ancient foe. N Engl J Med 363(1):87–89. doi:10.1056/NEJMcibr1004371

Howard JC, Hunn JP, Steinfeldt T (2011) The IRG protein-based resistance mechanism in mice and its relation to virulence in Toxoplasma gondii. Curr Opin Microbiol 14(4):414–421. doi:10.1016/j.mib.2011.07.002

Hunter CA, Sibley LD (2012) Modulation of innate immunity by Toxoplasma gondii virulence effectors. Nat Rev Microbiol 10(11):766–778. doi:10.1038/nrmicro2858

Hurles M (2004) Gene duplication: the genomic trade in spare parts. PLoS Biol 2(7):E206. doi:10.1371/journal.pbio.0020206

Imamura R, Wang Y, Kinoshita T, Suzuki M, Noda T, Sagara J, Taniguchi S, Okamoto H, Suda T (2010) Anti-inflammatory activity of PYNOD and its mechanism in humans and mice. J Immunol 184(10):5874–5884. doi:10.4049/jimmunol.0900779

Irvine KM, Andrews MR, Fernandez-Rojo MA, Schroder K, Burns CJ, Su S, Wilks AF, Parton RG, Hume DA, Sweet MJ (2009) Colony-stimulating factor-1 (CSF-1) delivers a proatherogenic signal to human macrophages. J Leukoc Biol 85(2):278–288. doi:10.1189/jlb.0808497

Isaacs A, Lindenmann J (1987) Virus interference. I. The interferon. By A. Isaacs and J. Lindenmann, 1957. J Interferon Res 7(5):429–438

J S (1966) The laboratory mouse. Biology of the Laboratory Mouse, New York

Janeway CA Jr (1989) Approaching the asymptote? Evolution and revolution in immunology. Cold Spring Harb Symp Quant Biol 54(Pt 1):1–13

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337(6096):816–821. doi:10.1126/science.1225829

Jurk M, Heil F, Vollmer J, Schetter C, Krieg AM, Wagner H, Lipford G, Bauer S (2002) Human TLR7 or TLR8 independently confer responsiveness to the antiviral compound R-848. Nat Immunol 3(6):499. doi:10.1038/ni0602-499

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20(10):1313–1326. doi:10.1101/gr.101386.109

Kamijo R, Harada H, Matsuyama T, Bosland M, Gerecitano J, Shapiro D, Le J, Koh SI, Kimura T, Green SJ et al (1994) Requirement for transcription factor IRF-1 in NO synthase induction in macrophages. Science 263(5153):1612–1615

Kanazawa N, Tashiro K, Miyachi Y (2004) Signaling and immune regulatory role of the dendritic cell immunoreceptor (DCIR) family lectins: DCIR, DCAR, dectin-2 and BDCA-2. Immunobiology 209(1–2):179–190. doi:10.1016/j.imbio.2004.03.004

Kanneganti TD, Lamkanfi M, Nunez G (2007) Intracellular NOD-like receptors in host defense and disease. Immunity 27(4):549–559

Kapetanovic R, Cavaillon JM (2007) Early events in innate immunity in the recognition of microbial pathogens. Expert opinion on biological therapy 7(6):907–918

Kapetanovic R, Fairbairn L, Downing A, Beraldi D, Sester DP, Freeman TC, Tuggle CK, Archibald AL, Hume DA (2013) The impact of breed and tissue compartment on the response of pig macrophages to lipopolysaccharide. BMC Genom 14(1):581. doi:10.1186/1471-2164-14-581

Kato H, Sato S, Yoneyama M, Yamamoto M, Uematsu S, Matsui K, Tsujimura T, Takeda K, Fujita T, Takeuchi O, Akira S (2005) Cell type-specific involvement of RIG-I in antiviral response. Immunity 23(1):19–28. doi:10.1016/j.immuni.2005.04.010

Kawai T, Akira S (2007) TLR signaling. Semin Immunol 19(1):24–32

Kawasaki K, Gomi K, Nishijima M (2001) Cutting edge: Gln22 of mouse MD-2 is essential for species-specific lipopolysaccharide mimetic action of taxol. J Immunol 166(1):11–14

Kingeter LM, Lin X (2012) C-type lectin receptor-induced NF-kappaB activation in innate immune and inflammatory responses. Cell Mol Immunol 9(2):105–112. doi:10.1038/cmi.2011.58

Klein J (2001) George Snell's first foray into the unexplored territory of the major histocompatibility complex. Genetics 159(2):435–439

Knirel YA, Lindner B, Vinogradov EV, Kocharova NA, Senchenkova SN, Shaikhutdinova RZ, Dentovskaya SV, Fursova NK, Bakhteeva IV, Titareva GM, Balakhonov SV, Holst O, Gremyakova TA, Pier GB, Anisimov AP (2005) Temperature-dependent variations and intraspecies diversity of the structure of the lipopolysaccharide of Yersinia pestis. Biochemistry 44(5):1731–1743. doi:10.1021/bi048430f

Koblansky AA, Jankovic D, Oh H, Hieny S, Sungnak W, Mathur R, Hayden MS, Akira S, Sher A, Ghosh S (2013) Recognition of profilin by Toll-like receptor 12 is critical for host resistance to Toxoplasma gondii. Immunity 38(1):119–130. doi:10.1016/j.immuni.2012.09.016

Koller BH, Hagemann LJ, Doetschman T, Hagaman JR, Huang S, Williams PJ, First NL, Maeda N, Smithies O (1989) Germ-line transmission of a planned alteration made in a hypoxanthine phosphoribosyltransferase gene by homologous recombination in embryonic stem cells. Proc Natl Acad Sci USA 86(22):8927–8931

Kollias G, Papadaki P, Apparailly F, Vervoordeldonk MJ, Holmdahl R, Baumans V, Desaintes C, Di Santo J, Distler J, Garside P, Hegen M, Huizinga TW, Jungel A, Klareskog L, McInnes I, Ragoussis I, Schett G, Hart B, Tak PP, Toes R, van den Berg W, Wurst W, Gay S (2011) Animal models for arthritis: innovative tools for prevention and treatment. Ann Rheum Dis 70(8):1357–1362. doi:10.1136/ard.2010.148551

Kotenko SV, Gallagher G, Baurin VV, Lewis-Antes A, Shen M, Shah NK, Langer JA, Sheikh F, Dickensheets H, Donnelly RP (2003) IFN-lambdas mediate antiviral protection through a distinct class II cytokine receptor complex. Nat Immunol 4(1):69–77. doi:10.1038/ni875

Kurt-Jones EA, Popova L, Kwinn L, Haynes LM, Jones LP, Tripp RA, Walsh EE, Freeman MW, Golenbock DT, Anderson LJ, Finberg RW (2000) Pattern recognition receptors TLR4 and CD14 mediate response to respiratory syncytial virus. Nat Immunol 1(5):398–401

Kwon DS, Gregorio G, Bitton N, Hendrickson WA, Littman DR (2002) DC-SIGN-mediated internalization of HIV is required for trans-enhancement of T cell infection. Immunity 16(1):135–144

Laayouni H, Oosting M, Luisi P, Ioana M, Alonso S, Ricano-Ponce I, Trynka G, Zhernakova A, Plantinga TS, Cheng SC, van der Meer JW, Popp R, Sood A, Thelma BK, Wijmenga C, Joosten LA, Bertranpetit J, Netea MG (2014) Convergent evolution in European and Rroma populations reveals pressure exerted by plague on Toll-like receptors. Proc Natl Acad Sci USA 111(7):2668–2673. doi:10.1073/pnas.1317723111

Lambert ND, Langfitt DM, Nilles ML, Bradley DS (2011) Resistance to Yersinia pestis infection decreases with age in B10.T(6R) mice. Infect Immun 79(11):4438–4446. doi:10.1128/IAI.05267-11

Leaman DW, Roberts RM (1992) Genes for the trophoblast interferons in sheep, goat, and musk ox and distribution of related genes among mammals. J Interferon Res 12(1):1–11

Lech M, Avila-Ferrufino A, Skuginna V, Susanti HE, Anders HJ (2010) Quantitative expression of RIG-like helicase, NOD-like receptor and inflammasome-related mRNAs in humans and mice. Int Immunol 22(9):717–728. doi:10.1093/intimm/dxq058

Lech M, Susanti HE, Rommele C, Grobmayr R, Gunthner R, Anders HJ (2012) Quantitative expression of C-type lectin receptors in humans and mice. Int J Mol Sci 13(8):10113–10131. doi:10.3390/ijms130810113

Lee JS, Campbell HD, Kozak CA, Young IG (1989) The IL-4 and IL-5 genes are closely linked and are part of a cytokine gene cluster on mouse chromosome 11. Somat Cell Mol Genet 15(2):143–152

Lee SM, Kok KH, Jaume M, Cheung TK, Yip TF, Lai JC, Guan Y, Webster RG, Jin DY, Peiris JS (2014) Toll-like receptor 10 is involved in induction of innate immune responses to influenza virus infection. Proc Natl Acad Sci USA 111(10):3793–3798. doi:10.1073/pnas.1324266111

Lehrer RI, Cole AM, Selsted ME (2012) theta-Defensins: cyclic peptides with endless potential. J Biol Chem 287(32):27014–27019. doi:10.1074/jbc.R112.346098

Lehrer RI, Lu W (2012) alpha-Defensins in human innate immunity. Immunol Rev 245(1):84–112. doi:10.1111/j.1600-065X.2011.01082.x

Lehrer RI, Szklarek D, Barton A, Ganz T, Hamann KJ, Gleich GJ (1989) Antibacterial properties of eosinophil major basic protein and eosinophil cationic protein. J Immunol 142(12):4428–4434

Lei B, Yao G (2006) Spectral attenuation of the mouse, rat, pig and human lenses from wavelengths 360 nm to 1020 nm. Exp Eye Res 83(3):610–614. doi:10.1016/j.exer.2006.02.013 S0014-4835(06)00182-5 [pii]

Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA (1996) The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. Cell 86(6):973–983

Li G, Zhang J, Sun Y, Wang H, Wang Y (2009) The evolutionarily dynamic IFN-inducible GTPase proteins play conserved immune functions in vertebrates and cephalochordates. Mol Biol Evol 26(7):1619–1630. doi:10.1093/molbev/msp074

Li R, Paul A, Ko KW, Sheldon M, Rich BE, Terashima T, Dieker C, Cormier S, Li L, Nour EA, Chan L, Oka K (2012) Interleukin-7 induces recruitment of monocytes/macrophages to endothelium. Eur Heart J 33(24):3114–3123. doi:10.1093/eurheartj/ehr245

Lilue J, Muller UB, Steinfeldt T, Howard JC (2013) Reciprocal virulence and resistance polymorphism in the relationship between Toxoplasma gondii and the house mouse. eLife 2:e01298. doi:10.7554/eLife.01298

Liu PT, Stenger S, Li H, Wenzel L, Tan BH, Krutzik SR, Ochoa MT, Schauber J, Wu K, Meinken C, Kamen DL, Wagner M, Bals R, Steinmeyer A, Zugel U, Gallo RL, Eisenberg D, Hewison M, Hollis BW, Adams JS, Bloom BR, Modlin RL (2006) Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response. Science 311(5768):1770–1773. doi:10.1126/science.1123933

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155

MacMicking JD, North RJ, LaCourse R, Mudgett JS, Shah SK, Nathan CF (1997) Identification of nitric oxide synthase as a protective locus against tuberculosis. Proc Natl Acad Sci USA 94(10):5243–5248

Magalhaes JG, Philpott DJ, Nahori MA, Jehanno M, Fritz J, Le Bourhis L, Viala J, Hugot JP, Giovannini M, Bertin J, Lepoivre M, Mengin-Lecreulx D, Sansonetti PJ, Girardin SE (2005) Murine Nod1 but not its human orthologue mediates innate immune detection of tracheal cytotoxin. EMBO Rep 6(12):1201–1207. doi:10.1038/sj.embor.7400552

Maloney B, Ge YW, Alley GM, Lahiri DK (2007) Important differences between human and mouse APOE gene promoters: limitation of mouse APOE model in studying Alzheimer's disease. J Neurochem 103(3):1237–1257. doi:10.1111/j.1471-4159.2007.04831.x

Marshall AS, Willment JA, Lin HH, Williams DL, Gordon S, Brown GD (2004) Identification and characterization of a novel human myeloid inhibitory C-type lectin-like receptor (MICL) that is predominantly expressed on granulocytes and monocytes. J Biol Chem 279(15):14792–14802. doi:10.1074/jbc.M313127200

Martens S, Sabel K, Lange R, Uthaiah R, Wolf E, Howard JC (2004) Mechanisms regulating the positioning of mouse p47 resistance GTPases LRG-47 and IIGP1 on cellular membranes: retargeting to plasma membrane induced by phagocytosis. J Immunol 173(4):2594–2606

Martinez FO, Gordon S, Locati M, Mantovani A (2006) Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression. J Immunol 177(10):7303–7311

Martinez J, Huang X, Yang Y (2010) Toll-like receptor 8-mediated activation of murine plasmacytoid dendritic cells by vaccinia viral DNA. Proc Natl Acad Sci USA 107(14):6442–6447. doi:10.1073/pnas.0913291107

Matsuguchi T, Takagi K, Musikacharoen T, Yoshikai Y (2000) Gene expressions of lipopolysaccharide receptors, toll-like receptors 2 and 4, are differently regulated in mouse T lymphocytes. Blood 95(4):1378–1385

Matzinger P (1994) Tolerance, danger, and the extended family. Annu Rev Immunol 12:991–1045

Maxwell AI, Morrison GM, Dorin JR (2003) Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. Mol Immunol 40(7):413–421

Mazzucchelli RI, Riva A, Durum SK (2012) The human IL-7 receptor gene: deletions, polymorphisms and mutations. Semin Immunol 24(3):225–230. doi:10.1016/j.smim.2012.02.007

McCollister BD, Bourret TJ, Gill R, Jones-Carson J, Vazquez-Torres A (2005) Repression of SPI2 transcription by nitric oxide-producing, IFNgamma-activated macrophages promotes maturation of Salmonella phagosomes. J Exp Med 202(5):625–635. doi:10.1084/jem.20050246

Medzhitov R, Preston-Hurlburt P, Janeway CA Jr (1997) A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. Nature 388(6640):394–397

Mestas J, Hughes CC (2004) Of mice and not men: differences between mouse and human immunology. J Immunol 172(5):2731–2738

Moser B, Wolf M, Walz A, Loetscher P (2004) Chemokines: multiple levels of leukocyte migration control. Trends Immunol 25(2):75–84. doi:10.1016/j.it.2003.12.005

Munford RS (2010) Murine responses to endotoxin: another dirty little secret? J Infect Dis 201(2):175–177. doi:10.1086/649558

Munk C, Wei G, Yang OO, Waring AJ, Wang W, Hong T, Lehrer RI, Landau NR, Cole AM (2003) The theta-defensin, retrocyclin, inhibits HIV-1 entry. AIDS Res Hum Retroviruses 19(10):875–881. doi:10.1089/088922203322493049

Murphy PM (1993) Molecular mimicry and the generation of host defense protein diversity. Cell 72(6):823–826 0092-8674(93)90571-7 [pii]

Murray HW, Szuro-Sudol A, Wellner D, Oca MJ, Granger AM, Libby DM, Rothermel CD, Rubin BY (1989) Role of tryptophan degradation in respiratory burst-independent antimicrobial activity of gamma interferon-stimulated human macrophages. Infect Immun 57(3):845–849

Namen AE, Lupton S, Hjerrild K, Wignall J, Mochizuki DY, Schmierer A, Mosley B, March CJ, Urdal D, Gillis S (1988a) Stimulation of B-cell progenitors by cloned murine interleukin-7. Nature 333(6173):571–573. doi:10.1038/333571a0

Namen AE, Schmierer AE, March CJ, Overell RW, Park LS, Urdal DL, Mochizuki DY (1988b) B cell precursor growth-promoting activity. Purification and characterization of a growth factor active on lymphocyte precursors. J Exp Med 167(3):988–1002

Nelson DE, Virok DP, Wood H, Roshick C, Johnson RM, Whitmire WM, Crane DD, Steele-Mortimer O, Kari L, McClarty G, Caldwell HD (2005) Chlamydial IFN-gamma immune evasion is linked to host infection tropism. Proc Natl Acad Sci USA 102(30):10658–10663. doi:10.1073/pnas.0504198102

Nomiyama H, Mera A, Ohneda O, Miura R, Suda T, Yoshie O (2001) Organization of the chemokine genes in the human and mouse major clusters of CC and CXC chemokines: diversification between the two species. Genes Immun 2(2):110–113. doi:10.1038/sj.gene.6363742

Nonaka M, Kimura A (2006) Genomic view of the evolution of the complement system. Immunogenetics 58(9):701–713

O'Dell JR (2004) Therapeutic strategies for rheumatoid arthritis. N Engl J Med 350(25):2591–2602. doi:10.1056/NEJMra040226

Ohmori Y, Hamilton TA (2001) Requirement for STAT1 in LPS-induced gene expression in macrophages. J Leukoc Biol 69(4):598–604

Ohno S (1970) Evolution by gene duplication. Springer, New York

Oldenburg M, Kruger A, Ferstl R, Kaufmann A, Nees G, Sigmund A, Bathke B, Lauterbach H, Suter M, Dreher S, Koedel U, Akira S, Kawai T, Buer J, Wagner H, Bauer S, Hochrein H, Kirschning CJ (2012) TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. Science 337(6098):1111–1115. doi:10.1126/science.1220363

Oliveira AC, Peixoto JR, de Arruda LB, Campos MA, Gazzinelli RT, Golenbock DT, Akira S, Previato JO, Mendonca-Previato L, Nobrega A, Bellio M (2004) Expression of functional TLR4 confers proinflammatory responsiveness to Trypanosoma cruzi glycoinositol phospholipids and higher resistance to infection with T. cruzi. J Immunol 173(9):5688–5696

Oppenheim JJ, Yang D (2005) Alarmins: chemotactic activators of immune responses. Curr Opin Immunol 17(4):359–365

Oritani K, Medina KL, Tomiyama Y, Ishikawa J, Okajima Y, Ogawa M, Yokota T, Aoyama K, Takahashi I, Kincade PW, Matsuzawa Y (2000) Limitin: an interferon-like cytokine that preferentially influences B-lymphocyte precursors. Nat Med 6(6):659–666. doi:10.1038/76233

Orlandi PA, Klotz FW, Haynes JD (1992) A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of Plasmodium falciparum recognizes the terminal Neu5Ac(alpha 2-3)Gal- sequences of glycophorin A. J Cell Biol 116(4):901–909

Ostler N, Britzen-Laurent N, Liebl A, Naschberger E, Lochnit G, Ostler M, Forster F, Kunzelmann P, Ince S, Supper V, Praefcke GJ, Schubert DW, Stockinger H, Herrmann C, Sturzl M (2014) Gamma interferon-induced guanylate binding protein 1 is a novel actin cytoskeleton remodeling factor. Mol Cell Biol 34(2):196–209. doi:10.1128/MCB.00664-13

Ouellette AJ, Selsted ME (1996) Paneth cell defensins: endogenous peptide components of intestinal host defense. FASEB J 10(11):1280–1289

Pabst R, Gehrke I (1990) Is the bronchus-associated lymphoid tissue (BALT) an integral structure of the lung in normal mammals, including humans? Am J Respir Cell Mol Biol 3(2):131–135. doi:10.1165/ajrcmb/3.2.131

Pal D, Dasgupta S, Kundu R, Maitra S, Das G, Mukhopadhyay S, Ray S, Majumdar SS, Bhattacharya S (2012) Fetuin-A acts as an endogenous ligand of TLR4 to promote lipid-induced insulin resistance. Nat Med 18(8):1279–1285. doi:10.1038/nm.2851

Pan J, Xia L, McEver RP (1998) Comparison of promoters for the murine and human P-selectin genes suggests species-specific and conserved mechanisms for transcriptional regulation in endothelial cells. J Biol Chem 273(16):10058–10067

Panitch HS, Hirsch RL, Haley AS, Johnson KP (1987) Exacerbations of multiple sclerosis in patients treated with gamma interferon. Lancet 1(8538):893–895

Park BS, Lee JO (2013) Recognition of lipopolysaccharide pattern by TLR4 complexes. Exp Mol Med 45:e66. doi:10.1038/emm.2013.97

Philpott DJ, Sorbara MT, Robertson SJ, Croitoru K, Girardin SE (2014) NOD proteins: regulators of inflammation in health and disease. Nat Rev Immunol 14(1):9–23. doi:10.1038/nri3565

Pickens SR, Chamberlain ND, Volin MV, Pope RM, Talarico NE, Mandelin AM 2nd, Shahrara S (2011) Characterization of interleukin-7 and interleukin-7 receptor in the pathogenesis of rheumatoid arthritis. Arthritis Rheum 63(10):2884–2893. doi:10.1002/art.30493

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(5396):2085–2088

Pugin J, Schurer-Maly CC, Leturcq D, Moriarty A, Ulevitch RJ, Tobias PS (1993) Lipopolysaccharide activation of human endothelial and epithelial cells is mediated by lipopolysaccharide-binding protein and soluble CD14. Proc Natl Acad Sci USA 90(7):2744–2748

Rast JP, Smith LC, Loza-Coll M, Hibino T, Litman GW (2006) Genomic insights into the immune system of the sea urchin. Science 314(5801):952–956. doi:10.1126/science.1134301

Regan T, Nally K, Carmody R, Houston A, Shanahan F, Macsharry J, Brint E (2013) Identification of TLR10 as a key mediator of the inflammatory response to Listeria monocytogenes in intestinal epithelial cells and macrophages. J Immunol 191(12):6084–6092. doi:10.4049/jimmunol.1203245

Rehli M (2002) Of mice and men: species variations of Toll-like receptor expression. Trends Immunol 23(8):375–378

Rensing SA (2014) Gene duplication as a driver of plant morphogenetic evolution. Curr Opin Plant Biol 17C:43–48. doi:10.1016/j.pbi.2013.11.002

Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, Hood LE, Aderem A (2005) The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci USA 102(27):9577–9582. doi:10.1073/pnas.0502272102

Roberts RM, Ezashi T, Rosenfeld CS, Ealy AD, Kubisch HM (2003) Evolution of the interferon tau genes and their promoters, and maternal-trophoblast interactions in control of their expression. Reprod Suppl 61:239–251

Roshick C, Wood H, Caldwell HD, McClarty G (2006) Comparison of gamma interferon-mediated antichlamydial defense mechanisms in human and mouse cells. Infect Immun 74(1):225–238. doi:10.1128/IAI.74.1.225-238.2006

Rovai LE, Herschman HR, Smith JB (1998) The murine neutrophil-chemoattractant chemokines LIX, KC, and MIP-2 have distinct induction kinetics, tissue distributions, and tissue-specific sensitivities to glucocorticoid regulation in endotoxemia. J Leukoc Biol 64(4):494–502

Ru H, Ni X, Zhao L, Crowley C, Ding W, Hung LW, Shaw N, Cheng G, Liu ZJ (2013) Structural basis for termination of AIM2-mediated signaling by p202. Cell Res 23(6):855–858. doi:10.1038/cr.2013.52

Rubino I, Coste A, Le Roy D, Roger T, Jaton K, Boeckh M, Monod M, Latge JP, Calandra T, Bochud PY (2012) Species-specific recognition of Aspergillus fumigatus by Toll-like receptor 1 and Toll-like receptor 6. J Infect Dis 205(6):944–954. doi:10.1093/infdis/jir882

Salzman NH, Ghosh D, Huttner KM, Paterson Y, Bevins CL (2003) Protection against enteric salmonellosis in transgenic mice expressing a human intestinal defensin. Nature 422(6931):522–526. doi:10.1038/nature01520

Sattler S, Ghadially H, Hofer E (2012) Evolution of the C-type lectin-like receptor genes of the DECTIN-1 cluster in the NK gene complex. Sci World J 2012:931386. doi:10.1100/2012/931386

Sattler S, Ghadially H, Reiche D, Karas I, Hofer E (2010) Evolutionary development and expression pattern of the myeloid lectin-like receptor gene family encoded within the NK gene complex. Scand J Immunol 72(4):309–318. doi:10.1111/j.1365-3083.2010.02433.x

Schaefer M, Reiling N, Fessler C, Stephani J, Taniuchi I, Hatam F, Yildirim AO, Fehrenbach H, Walter K, Ruland J, Wagner H, Ehlers S, Sparwasser T (2008) Decreased pathology and prolonged survival of human DC-SIGN transgenic mice during mycobacterial infection. J Immunol 180(10):6836–6845

Schmidt M, Raghavan B, Muller V, Vogl T, Fejer G, Tchaptchet S, Keck S, Kalis C, Nielsen PJ, Galanos C, Roth J, Skerra A, Martin SF, Freudenberg MA, Goebeler M (2010) Crucial role for human Toll-like receptor 4 in the development of contact allergy to nickel. Nat Immunol 11(9):814–819. doi:10.1038/ni.1919

Schroder K, Irvine KM, Taylor MS, Bokil NJ, Le Cao KA, Masterman KA, Labzin LI, Semple CA, Kapetanovic R, Fairbairn L, Akalin A, Faulkner GJ, Baillie JK, Gongora M, Daub CO, Kawaji H, McLachlan GJ, Goldman N, Grimmond SM, Carninci P, Suzuki H, Hayashizaki Y, Lenhard B, Hume DA, Sweet MJ (2012) Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. Proc Natl Acad Sci USA 109(16):E944–E953. doi:10.1073/pnas.1110156109

Schroder K, Tschopp J (2010) The inflammasomes. Cell 140(6):821–832. doi:10.1016/j.cell.2010.01.040

Schutte BC, Mitros JP, Bartlett JA, Walters JD, Jia HP, Welsh MJ, Casavant TL, McCray PB Jr (2002) Discovery of five conserved beta -defensin gene clusters using a computational search strategy. Proc Natl Acad Sci USA 99(4):2129–2133. doi:10.1073/pnas.042692699

Semple F, Dorin JR (2012) beta-Defensins: multifunctional modulators of infection, inflammation and more? J Innate Immun 4(4):337–348. doi:10.1159/000336619

Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, Richards DR, McDonald-Smith GP, Gao H, Hennessy L, Finnerty CC, Lopez CM, Honari S, Moore EE, Minei JP, Cuschieri J, Bankey PE, Johnson JL, Sperry J, Nathens AB, Billiar TR, West MA, Jeschke MG, Klein MB, Gamelli RL, Gibran NS, Brownstein BH, Miller-Graziano C, Calvano SE, Mason PH, Cobb JP, Rahme LG, Lowry SF, Maier RV, Moldawer LL, Herndon DN, Davis RW, Xiao W, Tompkins RG, Inflammation, Host Response to Injury LSCRP (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci USA 110(9):3507–3512. doi:10.1073/pnas.1222878110

Shanahan MT, Tanabe H, Ouellette AJ (2011) Strain-specific polymorphisms in Paneth cell alpha-defensins of C57BL/6 mice and evidence of vestigial myeloid alpha-defensin pseudogenes. Infect Immun 79(1):459–473. doi:10.1128/IAI.00996-10

Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, Feng T, Wakamatsu E, Benoist C, Koller D, Regev A, ImmGen C (2013) Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc Natl Acad Sci USA 110(8):2946–2951. doi:10.1073/pnas.1222738110

Shenderov K, Riteau N, Yip R, Mayer-Barber KD, Oland S, Hieny S, Fitzgerald P, Oberst A, Dillon CP, Green DR, Cerundolo V, Sher A (2014) Cutting edge: endoplasmic reticulum stress licenses macrophages to produce mature IL-1beta in response to TLR4 stimulation through a Caspase-8- and TRIF-Dependent pathway. J Immunol 192(5):2029–2033. doi:10.4049/jimmunol.1302549

Sheppard P, Kindsvogel W, Xu W, Henderson K, Schlutsmeyer S, Whitmore TE, Kuestner R, Garrigues U, Birks C, Roraback J, Ostrander C, Dong D, Shin J, Presnell S, Fox B, Haldeman B, Cooper E, Taft D, Gilbert T, Grant FJ, Tackett M, Krivan W, McKnight G, Clegg C, Foster D, Klucher KM (2003) IL-28, IL-29 and their class II cytokine receptor IL-28R. Nat Immunol 4(1):63–68. doi:10.1038/ni873

Shoham S, Huang C, Chen JM, Golenbock DT, Levitz SM (2001) Toll-like receptor 4 mediates intracellular signaling without TNF-alpha release in response to Cryptococcus neoformans polysaccharide capsule. J Immunol 166(7):4620–4626

Singh SB, Davis AS, Taylor GA, Deretic V (2006) Human IRGM induces autophagy to eliminate intracellular mycobacteria. Science 313(5792):1438–1441. doi:10.1126/science.1129577

Sitnicka E, Buza-Vidas N, Larsson S, Nygren JM, Liuba K, Jacobsen SE (2003) Human CD34 + hematopoietic stem cells capable of multilineage engrafting NOD/SCID mice express flt3: distinct flt3 and c-kit expression and response patterns on mouse and candidate human hematopoietic stem cells. Blood 102(3):881–886. doi:10.1182/blood-2002-06-1694

Sonnenburg JL, Altheide TK, Varki A (2004) A uniquely human consequence of domain-specific functional adaptation in a sialic acid-binding receptor. Glycobiology 14(4):339–346. doi:10.1093/glycob/cwh039

Specter S, Cimprich R, Friedman H, Chedid L (1978) Stimulation of an enhanced in vitro immune response by a synthetic adjuvant, muramyl dipeptide. J Immunol 120(2):487–491

Steeghs L, Keestra AM, van Mourik A, Uronen-Hansson H, van der Ley P, Callard R, Klein N, van Putten JP (2008) Differential activation of human and mouse Toll-like receptor 4 by the adjuvant candidate LpxL1 of Neisseria meningitidis. Infect Immun 76(8):3801–3807. doi:10.1128/IAI.00005-08

Su MY, Kuo CI, Chang CF, Chang CI (2013) Three-dimensional structure of human NLRP10/PYNOD pyrin domain reveals a homotypic interaction site distinct from its mouse homologue. PLoS ONE 8(7):e67843. doi:10.1371/journal.pone.0067843

Tada H, Nemoto E, Shimauchi H, Watanabe T, Mikami T, Matsumoto T, Ohno N, Tamura H, Shibata K, Akashi S, Miyake K, Sugawara S, Takada H (2002) Saccharomyces cerevisiae- and Candida albicans-derived mannan induced production of tumour necrosis factor alpha by human monocytes in a CD14- and Toll-like receptor 4-dependent manner. Microbiol Immunol 46(7):503–512

Takeuchi O, Kawai T, Muhlradt PF, Morr M, Radolf JD, Zychlinsky A, Takeda K, Akira S (2001) Discrimination of bacterial lipoproteins by Toll-like receptor 6. Int Immunol 13(7):933–940

Tanne A, Ma B, Boudou F, Tailleux L, Botella H, Badell E, Levillain F, Taylor ME, Drickamer K, Nigou J, Dobos KM, Puzo G, Vestweber D, Wild MK, Marcinko M, Sobieszczuk P, Stewart L, Lebus D, Gicquel B, Neyrolles O (2009) A murine DC-SIGN homologue contributes to early host defense against Mycobacterium tuberculosis. J Exp Med 206(10):2205–2220. doi:10.1084/jem.20090188

Tauber AI (2003) Metchnikoff and the phagocytosis theory. Nat Rev Mol Cell Biol 4(11):897–901. doi:10.1038/nrm1244

Teutsch SM, Booth DR, Bennetts BH, Heard RN, Stewart GJ (2003) Identification of 11 novel and common single nucleotide polymorphisms in the interleukin-7 receptor-alpha gene and their associations with multiple sclerosis. Eur J Hum Genet: EJHG 11(7):509–515. doi:10.1038/sj.ejhg.5200994

Thomas SR, Mohr D, Stocker R (1994) Nitric oxide inhibits indoleamine 2,3-dioxygenase activity in interferon-gamma primed mononuclear phagocytes. J Biol Chem 269(20):14457–14464

Thomma BP, Cammue BP, Thevissen K (2002) Plant defensins. Planta 216(2):193–202. doi:10.1007/s00425-002-0902-6

Ting JP, Lovering RC, Alnemri ES, Bertin J, Boss JM, Davis BK, Flavell RA, Girardin SE, Godzik A, Harton JA, Hoffman HM, Hugot JP, Inohara N, Mackenzie A, Maltais LJ, Nunez G, Ogura Y, Otten LA, Philpott D, Reed JC, Reith W, Schreiber S, Steimle V, Ward PA (2008) The NLR gene family: a standard nomenclature. Immunity 28(3):285–287. doi:10.1016/j.immuni.2008.02.005

Unterholzner L, Keating SE, Baran M, Horan KA, Jensen SB, Sharma S, Sirois CM, Jin T, Latz E, Xiao TS, Fitzgerald KA, Paludan SR, Bowie AG (2010) IFI16 is an innate immune sensor for intracellular DNA. Nat Immunol 11(11):997–1004. doi:10.1038/ni.1932

Uze G, Schreiber G, Piehler J, Pellegrini S (2007) The receptor of the type I interferon family. Curr Top Microbiol Immunol 316:71–95

Vallesi A, Giuli G, Ghiara P, Scapigliati G, Luporini P (1998) Structure-function relationships of pheromones of the ciliate Euplotes raikovi with mammalian growth factors: cross-reactivity between Er-1 and interleukin-2 systems. Exp Cell Res 241(1):253–259. doi:10.1006/excr.1998.4056

Varki A (2001) Loss of N-glycolylneuraminic acid in humans: Mechanisms, consequences, and implications for hominid evolution. American journal of physical anthropology Suppl 33:54–69

Vasseur E, Patin E, Laval G, Pajon S, Fornarino S, Crouau-Roy B, Quintana-Murci L (2011) The selective footprints of viral pressures at the human RIG-I-like receptor family. Hum Mol Genet 20(22):4462–4474. doi:10.1093/hmg/ddr377

Velez DR, Hulme WF, Myers JL, Weinberg JB, Levesque MC, Stryjewski ME, Abbate E, Estevan R, Patillo SG, Gilbert JR, Hamilton CD, Scott WK (2009) NOS2A, TLR4, and IFNGR1 interactions influence pulmonary tuberculosis susceptibility in African-Americans. Hum Genet 126(5):643–653. doi:10.1007/s00439-009-0713-y

Venkataraman N, Cole AL, Ruchala P, Waring AJ, Lehrer RI, Stuchlik O, Pohl J, Cole AM (2009) Reawakening retrocyclins: ancestral human defensins active against HIV-1. PLoS Biol 7(4):e95. doi:10.1371/journal.pbio.1000095

Verghese MW, Prince M, Snyderman R (1980) Genetic control of peripheral leukocyte response to endotoxin in mice. J Immunol 124(5):2468–2473

Viriyakosol S, Mathison JC, Tobias PS, Kirkland TN (2000) Structure-function analysis of CD14 as a soluble receptor for lipopolysaccharide. J Biol Chem 275(5):3144–3149

Visintin A, Halmen KA, Khan N, Monks BG, Golenbock DT, Lien E (2006) MD-2 expression is not required for cell surface targeting of Toll-like receptor 4 (TLR4). J Leukoc Biol 80(6):1584–1592. doi:10.1189/jlb.0606388

Visintin A, Mazzoni A, Spitzer JH, Wyllie DH, Dower SK, Segal DM (2001) Regulation of Toll-like receptors in human monocytes and dendritic cells. J Immunol 166(1):249–255

Walker AM, Roberts RM (2009) Characterization of the bovine type I IFN locus: rearrangements, expansions, and novel subfamilies. BMC Genom 10:187. doi:10.1186/1471-2164-10-187

Warren HS, Fitting C, Hoff E, Adib-Conquy M, Beasley-Topliffe L, Tesini B, Liang X, Valentine C, Hellman J, Hayden D, Cavaillon JM (2010a) Resilience to bacterial infection: difference between species could be due to proteins in serum. J Infect Dis 201(2):223–232. doi:10.1086/649557

Warren SE, Armstrong A, Hamilton MK, Mao DP, Leaf IA, Miao EA, Aderem A (2010b) Cutting edge: Cytosolic bacterial DNA activates the inflammasome via Aim2. J Immunol 185(2):818–821. doi:10.4049/jimmunol.1000724

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, Guigo 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(6915):520–562. doi:10.1038/nature01262 nature01262 [pii]

Webb DR (2014) Animal models of human disease: inflammation. Biochem Pharmacol 87(1):121–130. doi:10.1016/j.bcp.2013.06.014

Wehkamp J, Stange EF, Fellermann K (2009) Defensin-immunology in inflammatory bowel disease. Gastroenterol Clin Biol 33(Suppl 3):S137–S144. doi:10.1016/S0399-8320(09)73149-5

Wei XQ, Charles IG, Smith A, Ure J, Feng GJ, Huang FP, Xu D, Muller W, Moncada S, Liew FY (1995) Altered immune responses in mice lacking inducible nitric oxide synthase. Nature 375(6530):408–411. doi:10.1038/375408a0

Weinberg JB (1998) Nitric oxide production and nitric oxide synthase type 2 expression by human mononuclear phagocytes: a review. Mol Med 4(9):557–591

Wiens GD, Glenney GW (2011) Origin and evolution of TNF and TNF receptor superfamilies. Dev Comp Immunol 35(12):1324–1335. doi:10.1016/j.dci.2011.03.031

Woelk CH, Frost SD, Richman DD, Higley PE, Kosakovsky Pond SL (2007) Evolution of the interferon alpha gene family in eutherian mammals. Gene 397(1–2):38–50. doi:10.1016/j.gene.2007.03.018

Wong VW, Sorkin M, Glotzbach JP, Longaker MT, Gurtner GC (2011) Surgical approaches to create murine models of human wound healing. J Biomed Biotechnol 2011:969618. doi:10.1155/2011/969618

Wood-Charlson EM, Weis VM (2009) The diversity of C-type lectins in the genome of a basal metazoan. Nematostella vectensis. Dev Comp Immunol 33(8):881–889. doi:10.1016/j.dci.2009.01.008

Yarovinsky F, Zhang D, Andersen JF, Bannenberg GL, Serhan CN, Hayden MS, Hieny S, Sutterwala FS, Flavell RA, Ghosh S, Sher A (2005) TLR11 activation of dendritic cells by a protozoan profilin-like protein. Science 308(5728):1626–1629

Yin Q, Sester DP, Tian Y, Hsiao YS, Lu A, Cridland JA, Sagulenko V, Thygesen SJ, Choubey D, Hornung V, Walz T, Stacey KJ, Wu H (2013) Molecular mechanism for p202-mediated specific inhibition of AIM2 inflammasome activation. Cell reports 4(2):327–339. doi:10.1016/j.celrep.2013.06.024

Young JD, Peterson CG, Venge P, Cohn ZA (1986) Mechanism of membrane damage mediated by human eosinophil cationic protein. Nature 321(6070):613–616. doi:10.1038/321613a0

Young JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, Trask BJ (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. Hum Mol Genet 11(5):535–546

Yu M, Wang H, Ding A, Golenbock DT, Latz E, Czura CJ, Fenton MJ, Tracey KJ, Yang H (2006) HMGB1 signals through toll-like receptor (TLR) 4 and TLR2. Shock 26(2):174–179

Zanetti M (2004) Cathelicidins, multifunctional peptides of the innate immunity. J Leukoc Biol 75(1):39–48. doi:10.1189/jlb.0403147

Zelante T, Fallarino F, Bistoni F, Puccetti P, Romani L (2009) Indoleamine 2,3-dioxygenase in infection: the paradox of an evasive strategy that benefits the host. Microbes Infect 11(1):133–141. doi:10.1016/j.micinf.2008.10.007

Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat Biotechnol 29(2):149–153. doi:10.1038/nbt.1775

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95(7):3708–3713

Zhang X, Laubach VE, Alley EW, Edwards KA, Sherman PA, Russell SW, Murphy WJ (1996) Transcriptional basis for hyporesponsiveness of the human inducible nitric oxide synthase gene to lipopolysaccharide/interferon-gamma. J Leukoc Biol 59(4):575–585

Zhou H, Gu J, Lamont SJ, Gu X (2007) Evolutionary analysis for functional divergence of the toll-like receptor gene family and altered functional constraints. J Mol Evol 65(2):119–123. doi:10.1007/s00239-005-0008-4

Zlotnik A, Yoshie O (2000) Chemokines: a new classification system and their role in immunity. Immunity 12(2):121–127

Chapter 7
Evolutionary Genomics of Miniature Inverted-Repeat Transposable Elements (MITEs) in Plants

Jiongjiong Chen, Qun Hu, Chen Lu and Hanhui Kuang

Abstract More and more evidence has accumulated in the past 20 years suggesting that MITEs may have played important roles in plant gene and genome evolution. With a large number of plant genomes sequenced and the development of computational programs for de novo MITE identification, a massive number of MITEs have been identified from plant genomes. The number of MITEs in a genome varied dramatically among different plant species. There is significant correlation between the number of MITEs and genome size, though there are several prominent exceptions. Some MITE families have a high copy number in a genome, probably due to one or several rounds of amplification bursts. Different MITE families in the same genome may have experienced amplification burst at different times, suggesting that their amplifications were triggered by distinct environments (such as stress) or genetic events. However, very few MITEs in plant genomes are currently active. MITEs are often distributed in gene-rich regions, and may be inserted in genes' promoter regions or transcribed regions. They may affect (either upregulate or downregulate) the expression of nearby genes. MITEs may downregulate genes through small RNAs, which may be produced via NAT or double-stranded RNAs formed by transcribed MITE sequences. The presence/absence of MITEs as well as their potential effects on expression of nearby genes suggests that MITE may provide considerable physiological and phenotypic

J. Chen · Q. Hu · C. Lu · H. Kuang (✉)
Key Laboratory of Horticulture Biology, Ministry of Education,
and Department of Vegetable Crops, College of Horticulture and Forestry,
Huazhong Agricultural University, Wuhan 430070, People's Republic of China
e-mail: kuangfile@mail.hzau.edu.cn

J. Chen
e-mail: jjchen@mail.hzau.edu.cn

Q. Hu
e-mail: huqunwind@gmail.com

C. Lu
e-mail: chenlubio@gmail.com

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_7,

variations for a species. Important future studies on MITEs include the mechanisms of MITE activation and the effects of MITEs on gene and genome evolution.

Plant genomes harbor a large number of transposable elements (TE) (Feschotte et al. 2002; Tenaillon et al. 2010). Previously considered as junk and selfish DNAs, TEs are now known to play important roles in gene expression and genome evolution (Kazazian 2004; Fedoroff 2012; McCue and Slotkin 2012). According to their transposing intermediates, TEs are classified into two types, class I (RNA) and class II (DNA). The genome size of a plant species is largely determined by its class I TEs. However, a plant genome may also harbor a large number of short class II TEs, called miniature inverted-repeat transposable elements (MITEs).

7.1 MITE Features

MITEs are considered to be deletion derivatives of autonomous DNA transposons (Fig. 7.1). Like DNA transposons, MITEs usually have terminal inverted-repeat (TIR) sequences, which are flanked by target site duplication (TSD) sequences. Many MITE families have a large copy number in a genome. However, the definition of MITEs remains vague. By name, MITEs are supposed to be short (miniature), but the length of class II TEs has a continuous distribution, varying from dozens of bp to several kb (Fig. 7.2). The maximum length of MITEs has been defined inconsistently in the literature, varying from 500 to 800 bp (Naito et al. 2006; Lu et al. 2012; Han et al. 2013). No matter what maximum length was used to define MITEs, it was defined practically rather than scientifically. Nevertheless, a genome has low copy number of class II TEs that are larger than 800 bp (Fig. 7.2). Class II TE families with length >800 bp always have few elements in a genome, suggesting poor transposing capacity of large elements. On the other hand, some class II TEs may not have obvious TIR sequences. There is no reason (molecular or evolutionary) to consider these TEs with no TIRs as a unique group. For convenience, MITEs may refer to all class II TEs with short length (say <800 bp), regardless of their TIR, TSD, and copy number in a genome (Chen et al. 2013).

7.2 Identification of MITEs

The first MITE was discovered as a 128-bp insertion at the waxy locus of maize (Bureau and Wessler 1992), and subsequently many MITE families were identified through sequence analysis of insertions of 100–600 bp (Charrier et al. 1999; Bureau and Wessler 1994; Yang et al. 2001) With more and more information on MITE sequences and MITE features available, computer programs such as FINDMITE (Tu 2001), TRANSPO (Santiago et al. 2002), MUST (Chen et al.

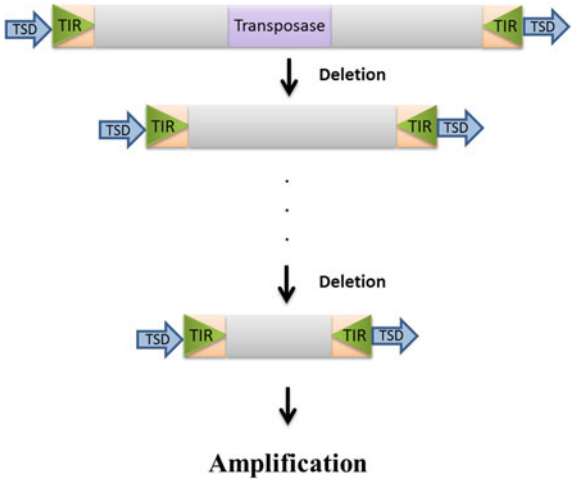


Fig. 7.1 MITE features. MITEs were derived from autonomous class II TEs after deletions of their central region including transposase-encoding sequences. Like autonomous class II TEs, MITEs usually have TIR sequences, flanked by short direst repeats (TSD). The TIR sequences may vary from a few to larger than 100 bp. The TSD sequences may vary from 2 to more than 10 bp. Note that some MITEs may not have perfect TIR sequences and/or lack of TSD sequences. After one or several rounds of internal deletions, the deletion derivatives (MITEs) may become active and amplify vigorously in a genome

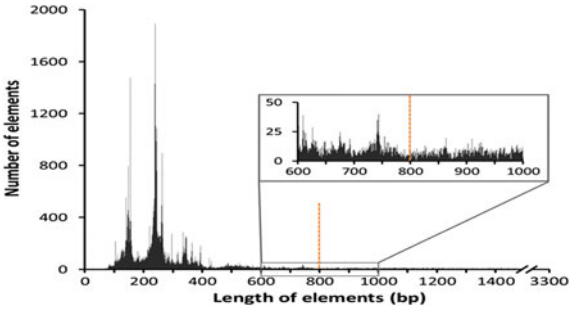


Fig. 7.2 Length distribution of all short class II TEs in the rice genome. Most class II TEs are 100–500 bp in length, and elements with length >800 bp are rare. Families with length >800 bp (marked by the *yellow line*) usually have few copies in a genome

2009), MITE-Hunter (Han and Wessler 2010), RSPB (Lu et al. 2012), and MITE Digger (Yang 2013) were developed to systematically identify MITEs from a database or genome sequences. Early programs had little success in MITE de novo detection. For example, FINDMITE and MUST generated high percentage of false positives (Han and Wessler 2010), while TRANSPO could not detect novel MITE families.

The programs MITE-Hunter, RSPB, and MITE Digger were developed recently for de novo identification of MITEs in a whole genome without any prerequisite MITE information. MITE-Hunter is a structure-based program and identifies repetitive sequences with TIR and TSD structures. It can identify a vast majority of

MITEs in a genome with low false-positive rate (4.4–8.3 %), and it can detect MITEs with a single copy in a genome. MITE-Hunter discovered 16 new MITE families in the rice genome besides 97.6 % of the annotated MITE sequences in the Repbase (Jurka et al. 2005). The program RSPB is BLASTN-based and it first identifies repetitive sequences with precise boundaries; then the identified families of repetitive sequences were further confirmed manually for MITE features such as TSD and TIR. Like MITE-Hunter, RSPB can detect the majority of MITEs in a genome with no prior information required. For example, 97.8 % rice MITE sequences deposited in Repbase were successfully identified by RSPB. RSPB and MITE-Hunter may identify MITE families that are missed by the others (Lu et al. 2012). In general, MITE-hunter runs fast and can detect MITE families with low copy number, while RSPB is more efficient than MITE-hunter in detecting MITEs with diverse TSD and TIR sequences. Output files of both programs are multiple sequence alignment files, which need laborious manual annotation. The program MITE Digger is approximately 3 and 15 times faster than MITE-Hunter and RSPB, respectively, and is fully automatic. However, it is difficult to judge whether the representative sequence from the output file of MITE Digger is a *bona fide* MITE, and it is difficult to categorize them since their TSD and terminal sequences may not be available in its output file. When the three programs MITE Digger, MITE-Hunter, and RSPB were applied together to a genome database, a vast majority of, if not all, MITE families are expected to be successfully identified. Recently, more than 2.3 million MITE sequences from 41 plant species were identified using the above three programs, and these MITE sequences can be searched and downloaded from a database (P-MITE) designed specifically for plant MITEs at http://pmite.hzau.edu.cn/django/mite/ (Chen et al. 2013).

7.3 Variation of MITE Types and MITE Copy Numbers Among Different Plant Species

Seven superfamilies of MITEs, including *Tc1/Mariner*, *PIF/Harbinger*, *CACTA*, *P, Novosib, hAT,* and *Mutator* have been identified in plant genomes (Kapitonov and Jurka 2008; Chen et al. 2013; Wicker et al. 2007). Elements from the *Tc1/Mariner* superfamily usually have 2 bp (TA) of TSD sequences, and their consensus TIR sequences are 5′-CTCCCTCC…GGAGGGAG-3′. MITEs of the *Tc1/Mariner* superfamily are usually short (Han et al. 2013). MITEs from *PIF/Harbinger* superfamily generate 3-bp TSDs (mainly TAA/TTA). The *CACTA* superfamily, which usually has TIR sequences starting with CACTA, also has 3-bp TSDs but with no nucleotide preference. To date, superfamilies *P* and *Novosib* have only been found in lower plant alga *Chlamydomonas reinhardtii* (Jurka et al. 2005). *P* elements create TSDs of 7–8 bp, while elements from the *Novosib* superfamily have 8-bp TSDs similar to that of superfamily *hAT*. The consensus of TIRs of the *hAT* superfamily is (T/C)A(A/G)NG with few exceptions (Rubin et al.

2001). Most MITEs from the *Mutator* superfamily produce TSDs of 9–10 bp. The TIR sequences of *Mutator* elements are usually long and highly divergent, and they start with G (few starting with C). MITEs with atypical TSD and TIR features would be difficult to be classified even though they may belong to one of above superfamilies by origin. Some MITE families are inserted only in microsatellites, but they do not have any other common features. It remains unclear whether they have a common ancestor or they should be classified as an independent superfamily. The MITEs inserted in microsatellite are collectively called *MiM* (for MITEs in microsatellite), and the *MiM* group represents the least frequent MITE group/superfamily in plant genomes (see below).

Analysis of MITEs in 41 plant genomes showed that the *Mutator* superfamily is the most prevalent, with an average of more than 20,000 copies per genome though it is absent in 8 of the 41 genomes investigated. In contrast to the *Mutator* superfamily, *MiM* is the least frequent group in higher plant genomes. Only ten higher plant genomes have *MiM*, with a total of 41,893 elements from 33 families. Among them, the strawberry genome contains 14 *MiM* families, while the other genomes have no more than 4 *MiM* families.

Plant genomes may vary considerably in total number of MITEs. The number of MITEs in lower plants is relatively few. For example, only 73 MITE elements are present in the genome of *Selaginella moellendorffii*, 83 in the genome of *Chlorella variabilis*, and no MITEs in the genome of *Cyanidioschyzon merolae*. Different higher plants may also vary in MITE copy number. For example, the papaya genome has only one MITE family (with 538 elements), while the apple genome has 180 MITE families with more than 200,000 elements. Closely related species may also vary dramatically in the number of MITEs. The number of MITEs in *Arabidopsis lyrata* is approximately five times as much as in *A. thaliana*, and the number of MITEs in watermelon is seven times as much as in melon (Chen et al. 2013).

Large plant genomes usually have a large number of MITEs, and there is a strong correlation coefficient between genome size and the number of MITEs in a genome (Fig. 7.3). However, several striking exceptions were exhibited. For instance, the papaya genome, with a genome size of 342 Mb, has only 538 elements, while the rice genome, which has a genome size similar to that of rice, contains 179,415 MITE elements.

7.4 Amplification of MITEs

Though MITEs and autonomous class II TEs are predicted to transpose via the same transposase, their copy numbers in a genome vary dramatically. An autonomous transposon family usually only has one to a handful of members, while a MITE family can have tens of thousands of copies (Fattash et al. 2013). It remains unknown why the deletion derivatives (MITEs) were much more successful in amplification than their full-length progenitors. One possibility is that the deletion

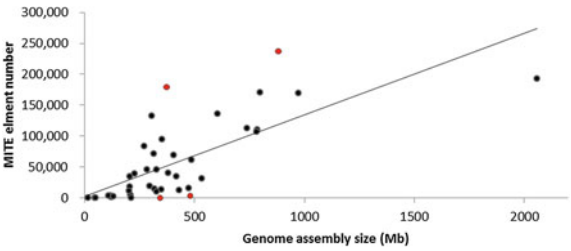


Fig. 7.3 The relationship between genome size and MITE copy number in a genome. Four prominent exceptions are marked in *red*

derivatives may form secondary structures that have better affinity with corresponding transposase and consequently facilitate their transpositions. Alternatively, MITEs may lack the regions with DNA modifications such as methylation that silences transposition of autonomous TEs. Interestingly, the length and structure of MITE families with high copy number are highly conserved within the same superfamily. For example, most MITE families from the *Tc1/Mariner* superfamily have length from 100 to 300 bp, though they were originated independently (i.e., the deletion events occurred independently) (Han et al. 2013). Furthermore, all MITE families of the *Tc1/Mariner* superfamily are mainly composed of TIR sequences, indicating that they were derived consistently from symmetric deletions of the middle of corresponding autonomous *Tc1/Mariner* elements. The conservation of length and structure of MITEs in the same superfamily in different plant species suggests that they were likely activated by homologous transposases.

However, shortened class II TEs do not guarantee efficient amplifications in a genome. Several MITE families were found to have only a few copies in a genome, in contrast to thousands of copies in other MITE families of the same superfamily (Chen et al. 2013). There is no evidence suggesting that low copy MITEs were generated recently since no highly similar autonomous TEs were found. Elements from large MITE families are often evenly related to each other, as evidenced by similar pairwise nucleotide identities among members (Lu et al. 2012). Such distribution of pairwise nucleotide identity suggests that a MITE family was expanded within a short period (i.e., with amplification burst) (Lu et al. 2012). Some MITE families have had two or more rounds of amplification bursts. The timing of amplification bursts varied among different MITE families, suggesting that there were no universal factors/events activating all or many MITE families simultaneously (see below).

7.5 Currently Active MITEs

Very few MITE families or full-length transposons are currently active in plant genomes (Jiang et al. 2003; Momose et al. 2010; Shirasawa et al. 2012; Fattash et al. 2013). The first confirmed active MITE was *mPing*. Sequence analysis first

predicted that the *mPing* family was active since it has many identical elements in rice genome (Jiang et al. 2003). Their transposability was experimentally verified using long-term cell culture (Jiang et al. 2003), anther culture (Kikuchi et al. 2003), and hybridization (Shan et al. 2005) and hydrostatic pressurization (Lin et al. 2006). Recently, another MITE family (*mGings*) was found to actively transpose in rice seedlings and plantlets regenerated from anther-derived calli (Dong et al. 2012). The *mGing* was first identified as a 146-bp insertion in the fourth intron of rice gene *LOC_Os01g04720* in one cultivar while absent in another. In other words, it was first identified simply as a MITE with presence/absence polymorphism among rice cultivars. Surprisingly, this MITE can be activated by γ-irradiation and tissue culture (Dong et al. 2012), considering that it has no other identical elements in rice genome before the stress treatments. It remains unknown whether other MITE families, besides *mPing* and *mGing* in the rice genome, can be activated by stresses such as γ-irradiation and tissue culture. If yes, it will be interesting to investigate the common features of these MITE families that can be activated by the same type of stress.

A purple-skinned somaclonal variant was obtained from a protoplast culture of a red-skinned potato (Momose et al. 2010). The purple skin was due to the gain of function of flavonoid 3′,5′-hydroxylase, which was silenced by an insertion of a *Stowaway* MITE in its first exon. Protoplast culture activated the MITE, the MITE transposed away from the gene, and then the function of gene resumed. Therefore, like the *mPing* family in rice, the *Stowaway* MITE in potato can be activated under particular conditions such as tissue culture. Two peanut mutants derived from diethyl sulfate (DES) mutagenesis have high-oleate phenotype, which was caused by mutations in the gene *ahFAD2A* (Shirasawa et al. 2012). The mutations were caused by insertions of MITEs. Surprisingly, the inactivation of the gene was caused by independent insertions of the same MITE, since the insertions in the two mutants were at different positions of the gene (Patel et al. 2004). It may suggest that this MITE in peanut was very active after DES treatment. Alternatively (but not exclusively), it may suggest that other TEs (including other MITEs) in peanut genome were rarely activated by DES treatment.

In summary, the transposition of several MITEs was observed in plant genomes. However, all of them were activated by stresses such as tissue culture. A systemic study on the activity of all MITEs in a genome under different stresses will provide valuable information on how MITEs are activated. Such knowledge will help us to understand how some MITE families were efficiently activated during a short period of evolutionary history and accumulated to high copy number in a genome. With more and more plant genome sequences available, it will be trivial to identify MITEs with many identical elements in a genome. MITE families with many identical elements are potentially active and a comprehensive study of these potential active MITEs will provide insights on how MITEs are activated and evolved.

7.6 MITEs Are Frequently Transcribed with Genes

MITEs are frequently inserted into gene-rich regions, and therefore have a high chance to alter the structure of genes and/or affect their expressions. If inserted into genes, MITE sequences may provide transcription start sites, poly(A) signals, exons, and splice junctions for genes (Oki et al. 2008). The prominent effects of MITEs on genes may lie on their frequent transcription with genes. More than 7,000 MITEs in rice cultivar Nipponbare are transcribed, and approximately half of them are present in transcripts with coding sequences (Lu et al. 2012). MITE sequences were also frequently found in the EST database of other plant species (Kuang et al. 2009).

7.7 MITE-Derived Small RNAs

The TIR sequences when transcribed may form double-stranded RNA (dsRNA) and then generate small RNAs. Indeed, approximately a quarter of small RNAs in rice were generated by MITE sequences (Lu et al. 2012). MITE-derived small RNAs, primarily 24 nucleotides in length, were also found in Solanaceae species (Kuang et al. 2009).

The MITE-derived small RNAs may be produced through microRNAs (miRNAs) or small interfering RNAs (siRNAs) biogenesis pathway. The long TIR sequences of some MITEs, when transcribed, can form a stem-loop (hair-pin) structure similar to that of miRNA genes. In fact, MITEs have been suggested as the ancestors of some miRNA genes (Piriyapongsa and Jordan 2007, 2008). However, most MITEs have short or no TIR sequences and therefore small RNAs from such MITE sequences are unlikely generated through the miRNA biogenesis pathway. Homologous MITEs may be inserted into genes in alternative orientations, forming natural sense/antisense transcripts (NAT) if co-expressed. More than 1,000 pairs of NAT were predicted in the rice genome (Lu et al. 2012). It is well known that the NAT can produce small RNAs through the siRNA biogenesis pathway (Borsani et al. 2005; Zhang et al. 2012). The factors determining the abundance of MITE-derived small RNAs remain unclear. The expression level of corresponding transcripts and its secondary structure might play important roles in the efficiency of small RNA production. It was shown that the main parts to produce small RNAs vary dramatically among different MITE families. For some MITE families, small RNAs are predominantly generated from their terminals, while they are mainly from central parts for other MITE families (Lu et al. 2012). The regions with abundant MITE-derived small RNAs usually have high G/C content (Zhang and Kuang, unpublished data), consistent with previous results that small RNAs have higher GC content than genome-wide average (Ho et al. 2007).

7.8 Effects of MITEs on Gene Expression

MITEs that are frequently inserted into genes or near genes may affect the expression of nearby genes in several ways. MITEs may contain some regulatory motifs and can regulate gene expression if they are inserted in the promoter region. MITE *Kiddo* in rice was shown to upregulate the expression of gene *Ubiquitin2* when inserted in its promoter region (Yang et al. 2005). The insertions of *mPing* MITE family into the upstream of rice genes had a modest effect on gene transcription (Naito et al. 2009). On average, genes with MITE insertions in rice have higher expression than those with no MITE insertions (Han et al. 2013). However, opposite conclusions were drawn by a parallel study, in which MITEs are more likely associated with genes of low expression (Lu et al. 2012). Comparison between genes with MITE insertion in one cultivar and no MITE insertion in another cultivar led to inconclusive results: upregulation, downregulation and no effects were almost equally encountered for genes with MITE insertion. Therefore, MITEs may upregulate gene expression in some cases while downregulate gene expression in other cases. The downregulation is very likely via MITE-derived small RNAs. MITE-derived small RNAs may cause modification (such as methylation) of MITE sequences as well as their nearby genes. MITEs near genes are maintained in a species only when their effects (either silencing or upregulating) increase species fitness. It is possible that many MITEs inserted near genes have been purged since silencing of TEs near genes by small RNAs may have deleterious effects on neighboring gene expression, and TEs targeted by small RNAs tend to be more likely to be deleted from plant genomes (Hollister and Gaut 2009; Wang et al. 2013).

7.9 Evolution of MITEs

There is much evidence showing that MITEs are deletion derivatives of autonomous class II TEs (Zhang et al. 2001; Feschotte et al. 2002; Jiang et al. 2003; Menzel et al. 2006; Dong et al. 2012). The deletions within a TE might have happened randomly, but the deletion size and position may determine their subsequent amplification efficiency. Only when the deletion remnants have a preferred length and structure could they become successful MITEs. The newly developed MITEs, such as the *mPing*, may be able to transpose right after birth in normal conditions, while some MITEs can transpose only under certain circumstances such as stresses. Alternatively, some MITEs might have been activated only after point mutations at critical sites (Chen and Kuang unpublished data).

Homologous MITEs were grouped into a family for convenience (Chen et al. 2013). However, elements from a "MITE family" may have been originated independently from different autonomous class II TEs. Similarly, homologous MITE families from two species might have been generated independently after speciation. Nevertheless, only two of the 338 MITE families in rice have homologous sequences

in the genome of *Brachypodium distachyon*, which was diverged from rice ~50 MYA (Kellogg 2001), suggesting that most of the recognizable MITEs were generated after speciation of rice and *B. distachyon* (Chen et al. 2012).

MITEs were ubiquitous in eukaryote genomes (Fattash et al. 2013; Yang 2013; Zhang et al. 2013), and should be present in the common ancestor of related species such as rice and *B. distachyon*. The paucity of homologous MITEs in the two species may suggest that MITE sequences with common ancestors are so diverged in the two species that they do not show significant sequence similarities any more. Under this hypothesis, most ancient MITE elements no longer have MITE features and are indistinguishable from other intergenic sequences. In contrast, some ancient MITEs still maintain the necessary features that have helped them transpose and amplify after speciation. For those ancient but extant MITEs, it is difficult to identify their progenitors since they have diverged for a long evolutionary time.

7.10 Diversity Generated by MITE Insertions

MITEs that have deleterious effects on neighboring gene expression will be ultimately deleted from plant genomes (Hollister and Gaut 2009; Wang et al. 2013). On the other hand, MITEs may be selected for and fixed in a species if they promote species fitness. If the presence or absence of a MITE insertion is selected for under different environments, presence/absence polymorphism of MITEs will prevail. Of course, presence/absence polymorphism can also be caused by genetic drift, particular for young MITE sites that have no effects on fitness. Comparison between rice cultivar Nipponbare and another rice cultivar 93–11 showed 14.8 % of MITEs exhibit presence/absence polymorphism (Chen et al. 2012). Considering that MITEs play important roles in the expression of nearby genes, the presence/absence polymorphism of MITEs may provide considerable phenotypic variations for a species.

Acknowledgments This research was supported by National Natural Science Foundation of China [grant no. 31300299 and 30921002].

References

Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005) Endogenous siRNAs derived from a pair of natural *cis*-antisense transcripts regulate salt tolerance in *Arabidopsis*. Cell 123(7):1279–1291

Bureau TE, Wessler SR (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4(10):1283–1294

Bureau TE, Wessler SR (1994) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6(6):907–916

Charrier B, Foucher F, Kondorosi E, d'Aubenton-Carafa Y, Thermes C, Kondorosi A, Ratet P (1999) Bigfoot. a new family of MITE elements characterized from the Medicago genus. Plant J 18(4):431–441

Chen J, Hu Q, Zhang Y, Lu C, Kuang H (2013) P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res. doi:10.1093/nar/gkt1000

Chen J, Lu C, Zhang Y, Kuang H (2012) Miniature inverted-repeat transposable elements (MITEs) in rice were originated and amplified predominantly after the divergence of Oryza and Brachypodium and contributed considerable diversity to the species. Mob Genet Elem 2(3):127–132

Chen Y, Zhou F, Li G, Xu Y (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. Gene 436(1–2):1–7

Dong HT, Zhang L, Zheng KL, Yao HG, Chen J, Yu FC, Yu XX, Mao BZ, Zhao D, Yao J, Li DB (2012) A *Gaijin-like* miniature inverted repeat transposable element is mobilized in rice during cell differentiation. BMC Genomics 13:135

Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P, Yang G (2013) Miniature inverted-repeat transposable elements: discovery, distribution, and activity. Genome 56(9):475–486

Fedoroff NV (2012) Presidential address. Transposable elements, epigenetics, and genome evolution. Science 338(6108):758–767

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5):329–341

Han Y, Qin S, Wessler SR (2013) Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics 14:71

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38(22):e199

Ho T, Wang H, Pallett D, Dalmay T (2007) Evidence for targeting common siRNA hotspots and GC preference by plant Dicer-like proteins. FEBS Lett 581(17):3267–3272

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19(8):1419–1428

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR (2003) An active DNA transposon family in rice. Nature 421(6919):163–167

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1–4):462–467

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9(5):411–412

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303(5664): 1626–1632

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol 125(3):1198–1205

Kikuchi K, Terauchi K, Wada M, Hirano HY (2003) The plant MITE *mPing* is mobilized in anther culture. Nature 421(6919):167–170

Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, Jiang J, Buell CR, Baker B (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. Genome Res 19(1):42–56

Lin X, Long L, Shan X, Zhang S, Shen S, Liu B (2006) In planta mobilization of *mPing* and its putative autonomous element *Pong* in rice by hydrostatic pressurization. J Exp Bot 57(10):2313–2323

Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H (2012) Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. Mol Biol Evol 29(3):1005–1017

McCue AD, Slotkin RK (2012) Transposable element small RNAs as regulators of gene expression. Trends Genet 28(12):616–623

Menzel G, Dechyeva D, Keller H, Lange C, Himmelbauer H, Schmidt T (2006) Mobilization and evolutionary history of miniature inverted-repeat transposable elements (MITEs) in *Beta vulgaris L.* Chromosome Res 14(8):831–844

Momose M, Abe Y, Ozeki Y (2010) Miniature inverted-repeat transposable elements of Stowaway are active in potato. Genetics 186(1):59–66

Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103(47):17620–17625

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461(7267):1130–1134

Oki N, Yano K, Okumoto Y, Tsukiyama T, Teraishi M, Tanisaka T (2008) A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. Genes Genet Syst 83(4):321–329

Patel M, Jung S, Moore K, Powell G, Ainsworth C, Abbott A (2004) High-oleate peanut mutants result from a MITE insertion into the *FAD2* gene. Theor Appl Genet 108(8):1492–1502

Piriyapongsa J, Jordan IK (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE 2(2):e203

Piriyapongsa J, Jordan IK (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. RNA 14(5):814–821

Rubin E, Lithwick G, Levy AA (2001) Structure and evolution of the *hAT* transposon superfamily. Genetics 158(3):949–957

Santiago N, Herraiz C, Goni JR, Messeguer X, Casacuberta JM (2002) Genome-wide analysis of the Emigrant family of MITEs of *Arabidopsis thaliana*. Mol Biol Evol 19(12):2285–2293

Shan X, Liu Z, Dong Z, Wang Y, Chen Y, Lin X, Long L, Han F, Dong Y, Liu B (2005) Mobilization of the active MITE transposons *mPing* and *Pong* in rice by introgression from wild rice (*Zizania latifolia Griseb.*). Mol Biol Evol 22(4):976–990

Shirasawa K, Hirakawa H, Tabata S, Hasegawa M, Kiyoshima H, Suzuki S, Sasamoto S, Watanabe A, Fujishiro T, Isobe S (2012) Characterization of active miniature inverted-repeat transposable elements in the peanut genome. Theor Appl Genet 124(8):1429–1438

Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends Plant Sci 15(8):471–478

Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc Natl Acad Sci USA 98(4):1699–1704

Wang X, Weigel D, Smith LM (2013) Transposon variants and their effects on gene expression in *Arabidopsis*. PLoS Genet 9(2):e1003255

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12):973–982

Yang G (2013) MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. BMC Bioinform 14:186

Yang G, Dong J, Chandrasekharan MB, Hall TC (2001) *Kiddo*, a new transposable element family closely associated with rice genes. Mol Genet Genomics 266(3):417–424

Yang G, Lee YH, Jiang Y, Shi X, Kertbundit S, Hall TC (2005) A two-edged role for the transposable element *Kiddo* in the rice ubiquitin2 promoter. Plant Cell 17(5):1559–1568

Zhang HH, Xu HE, Shen YH, Han MJ, Zhang Z (2013) The origin and evolution of six miniature inverted-repeat transposable elements in *Bombyx mori* and *Rhodnius prolixus*. Genome Biol Evol 5(11):2020–2031

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR (2001) *P* instability factor: an active maize transposon system associated with the amplification of *Tourist-like* MITEs and a new superfamily of transposases. Proc Natl Acad Sci USA 98(22):12572–12577

Zhang X, Xia J, Lii YE, Barrera-Figueroa BE, Zhou X, Gao S, Lu L, Niu D, Chen Z, Leung C, Wong T, Zhang H, Guo J, Li Y, Liu R, Liang W, Zhu JK, Zhang W, Jin H (2012) Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. Genome Biol 13(3):R20

Chapter 8
Horizontal Gene Transfer and the Role of Restriction-Modification Systems in Bacterial Population Dynamics

George Vernikos and Duccio Medini

Abstract Horizontal gene transfer (HGT) mediates non-vertical exchange of genetic elements thereby obfuscating the phylogenetic signal associated with vertically inherited mutations. Bacterial species exposed to significant HGT deviate from the clonal paradigm of a-sexual reproduction, towards pan-mictic admixtures. Intermediate population structures were also observed in which, despite high HGT rates, well-defined lineages coexist with a pan-mictic background. Different "forces" have been proposed to account for the containment of the HGT pan-mixing effect, including selection, fitness-related expansions and micro-epidemic evolution. Restriction-modification systems (RMSs) modulate the length of horizontally transferred DNA by selective cleavage of genetic material with heterologous methylation patterns. In a pan-genomic analysis of the *Neisseria meningitidis* bacterial species, sets of RMSs associated to specific lineages were shown to generate a differential barrier to DNA exchange, consistent with the inferred phylogeny.These data suggest that HGT, instead of being a "force" opposed to the emergence, persistence and global dissemination of consistent lineages, when modulated by RMSs can be the very cause of the intermediate population structures observed for the majority of pathogenic bacteria.

8.1 Horizontal Gene Transfer

Perhaps very few themes in the study of microbial evolution have been as contentious as Horizontal Gene Transfer (HGT) (Kurland 2000; Lawrence and Hendrickson 2003). HGT is defined as the transfer of genetic material between a

G. Vernikos
Novartis (Hellas) S.A.C.I., Athens, Greece

D. Medini (✉)
Novartis Vaccines Research, Siena, Italy
e-mail: duccio.medini@novartis.com

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_8,

donor and a recipient, in which no asexual (or sexual) reproduction is involved; the donor need not be physically present.

Early discussion on HGT came from Griffith, in a study focused on the ability of pneumococci to exchange genetic material through direct uptake of DNA from the environment (transformation) (Griffith 1928); later on Anderson and Syvanen (Anderson 1970; Syvanen 1985) discussed the concept of gene transfer across species boundaries.

HGT as a concept has fuelled very strong and ongoing debate about its impact, extent, gene, and host repertoire affected and frequency throughout the evolution of species (Kurland 2000; Lawrence and Hendrickson 2003). The controversy stems mainly from the fact that HGT is a counterintuitive concept that threatens to reject (Doolittle et al. 1996; Lawrence 2002; Gevers et al. 2005) the universality of a very fundamental biological concept, that of the biological species (Mayr 1942); furthermore it brings into question the Tree of Life (Darwin 1859), i.e., the representation of the phylogenetic history and evolution of species through a strictly bifurcating tree-like structure.

In terms of its impact, views range (Lawrence and Hendrickson 2003) from HGT being a valid but nonetheless rare mechanism of gene transfer with marginal impact on genome phylogeny (Kurland et al. 2003), to HGT being a major driving force that enables accelerated microbial evolution, often referred to as "evolution in quantum leaps" (Groisman and Ochman 1996); for example, two single-step events of HGT enabled Salmonella to evade successfully the host defense mechanisms and invade epithelial cells (Hacker et al. 1997). Supporters of the first view put forward the idea that the evolutionary history of a species can still be reliably represented through a bifurcating tree-like structure that reflects mainly the organismal phylogeny (Woese 2000; Daubin et al. 2003; Kurland et al. 2003; Lerat et al. 2005) since HGT frequency is not high enough to obscure the true phylogenetic signal of a given species. Supporters of the second opinion, however, believe that HGT can obfuscate the organismal phylogenetic signal to such an extent (i.e., mosaic genomes that contain genes with different histories) that the reliable representation of the organismal phylogeny violates the strictly bifurcating structure of the Tree of Life; instead reticulate, network-like structures can more reliably represent the true phylogenetic relationships between species that extensively exchange genetic material (Doolittle 1999b; Gogarten and Townsend 2005; Kunin et al. 2005).

For example, two distantly related species that have extensively exchanged genetic material with each other, now having mosaic genomes with patches of DNA with different histories, will probably map (wrongly) on very close branches on the phylogenetic tree, since their phylogenetic histor(y)ies are forced to fit in a strictly binary (i.e., either they belong to the same species or not) classification system. On the other hand, acknowledging that mosaicism is a valid genomic state, we can allow genomes to belong to more than one species at the same time (Doolittle 1999b); under a phylogenetic network representation the same two genomes will map correctly on their respective species/genera branches but their extensive genetic exchange will also be taken into account, represented through multiple branches connecting the two lineages.

It should be noted that similar results of genome mosaicism with patches of very similar DNA shared between very closely related taxa may also be attributed to genetic exchange via homologous recombination (Feil et al. 2001; Didelot et al. 2007). An example that illustrates the extent of viable genomic mosaicism, and at the same time questions the true boundaries of the biological species concept, comes from the model bacterial organism *Escherichia coli*; a three way comparison between the laboratory strain MG1655, the uropathogenic (UPEC) strain CFT073, and the enterohemorrhagic (EHEC) strain EDL933, shows that less than 40 % of their common gene pool is shared between those three strains, although their high sequence similarity places them under the same species (Welch et al. 2002).

At this point it may be useful to draw a parallel with quantum mechanics to discuss further the limitations of a binary classification system when describing complex biological processes. According to the classical Bohr model (Bohr 1913) of the atom, electrons (in our case genomes) are allowed to belong only to one of the well-defined orbits (in our case species) around the nucleus. Later on, however, the quantum mechanics theory (Schrödinger 1926) introduced a new, more realistic representation of the atom structure: the electrons surrounding the nucleus belong to a cloud (in our case phylogenetic network) of probable positions, rather than single well-defined orbits. The existence of the first atom model (in our case the tree of life) was due to our inability to study in a more detailed and realistic way the true structure of the atom (in our case the history of species); more sophisticated, nonbinary methods bring a more realistic view in our understanding and modeling of the history of species evolution (Fig. 8.1).

From the host point of view, the extent of HGT ranges from 0 % in *Buchnera aphidicola* (Tamas et al. 2002) to 24 % in *Thermotoga maritime* (Nelson et al. 1999); from the donor point of view, the extent of HGT might be up to 100 %, i.e., whole-genome transfer of a donor to a recipient cell (Dunning Hotopp et al. 2007).

Examples of HGT events exist in all three domains of life, i.e., bacteria (Baumler 1997; Lawrence and Ochman 1997), archaea (Gribaldo et al. 1999; Deppenmeier et al. 2002), and eukaryote (Dunning Hotopp et al. 2007), including humans, although the extent of HGT in the latter is not very well documented (Andersson et al. 2001; Stanhope et al. 2001).

In terms of gene repertoire, again HGT seems to affect a wide range of functional gene classes including genes encoding products involved in the translation machinery (e.g., aminoacyl-tRNA synthetases, ribosomal proteins) (Wolf et al. 1999; Brochier et al. 2000), ribosomal RNA (rRNA) genes (Nomura 1999; Yap et al. 1999), components of biosynthetic pathways (e.g., cytochrome c biogenesis system I and II) (Goldman and Kranz 1998) and major metabolic components (e.g., glyceraldehyde-3-phosphate dehydrogenase) (Doolittle et al. 1990); a good review on how HGT might have affected major metabolic pathways is given by Boucher et al. (2003).

Although in theory all genes can be horizontally exchanged, some functional classes (e.g., operational genes) may be more frequently transferred than others (e.g., informational) (Jain et al. 1999). Estimates of the actual frequency of HGT events in microbial genomes exist and suggest that HGT can be indeed a very

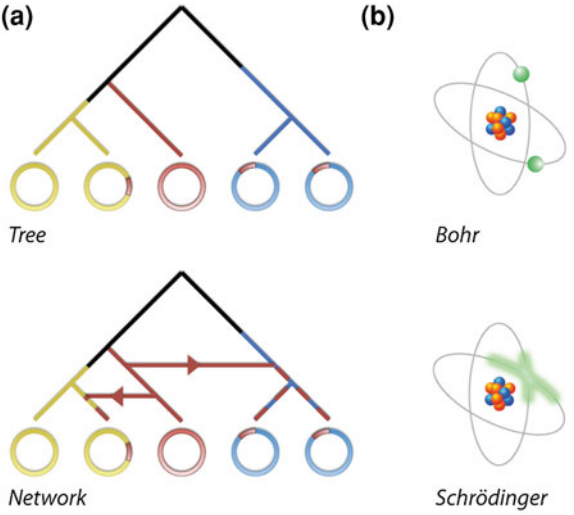


Fig. 8.1 Phylogenetic networks account for horizontal genetic flux. **a** An example of genome mosaicism and the limitation of a bifurcating, tree-based classification system (*top*) for a reliable representation of the true phylogenetic histories of lineages exposed to high rates of genetic flux, compared to a phylogenetic network (*bottom*), **b** Schematic comparison of the electron orbital representation under the Bohr (*top*) and the Schrödinger quantum theory (*bottom*)

frequent mechanism of gene transfer. Lawrence and Ochman (1997) studying the effects of HGT in *E. coli* and *S. enterica* estimated the HGT rate to be 31 kb per million years (Myr); this rate is close to the frequency of DNA being introduced by point mutations. Applying this rate of HGT, the two sister lineages were predicted to have each gained and lost over 3 Mb of alien DNA, since their divergence, approximately 100–140 million years (Myr) ago (Ochman and Wilson 1987; Doolittle et al. 1996).

Although horizontally acquired DNA enters a different, completely new genomic environment of a another host, the expression of horizontally acquired genes is not random or unrestrained; on the contrary the expression of alien DNA can be extremely sophisticated and fine-tuned. For example in Salmonella the quorum sensing mechanism that controls the cell population density directly affects the expression of genes that have been en block horizontally acquired under a single event (Choi et al. 2007). Similarly, SlyA, a virulence-related transcriptional regulator, participates in the regulation of another block of alien genes present in *S. enterica* (Linehan et al. 2005). A putative master regulator of the expression of horizontally acquired DNA has been recognized in enterobacteriaceae (Navarre et al. 2006): H-NS, a histone-like nucleoid structuring protein has been proposed to be responsible for selectively silencing horizontally acquired DNA of lower G+C% content relative the backbone composition of the host. It is worth noting that SlyA acts as an antagonist to H-NS, displacing the H-NS from promoter loci (Wyborn et al. 2004), adding one extra level of complexity to the regulatory network controlling the expression of alien DNA in microbial genomes.

There are three reported major mechanisms of HGT (Fig. 8.2), namely transformation (Griffith 1928), conjugation (Lederberg 1956), and transduction (Morse et al. 1956). A major difference between conjugation and the other two types of gene transfer, in terms of the donor and the recipient, is that in transduction and

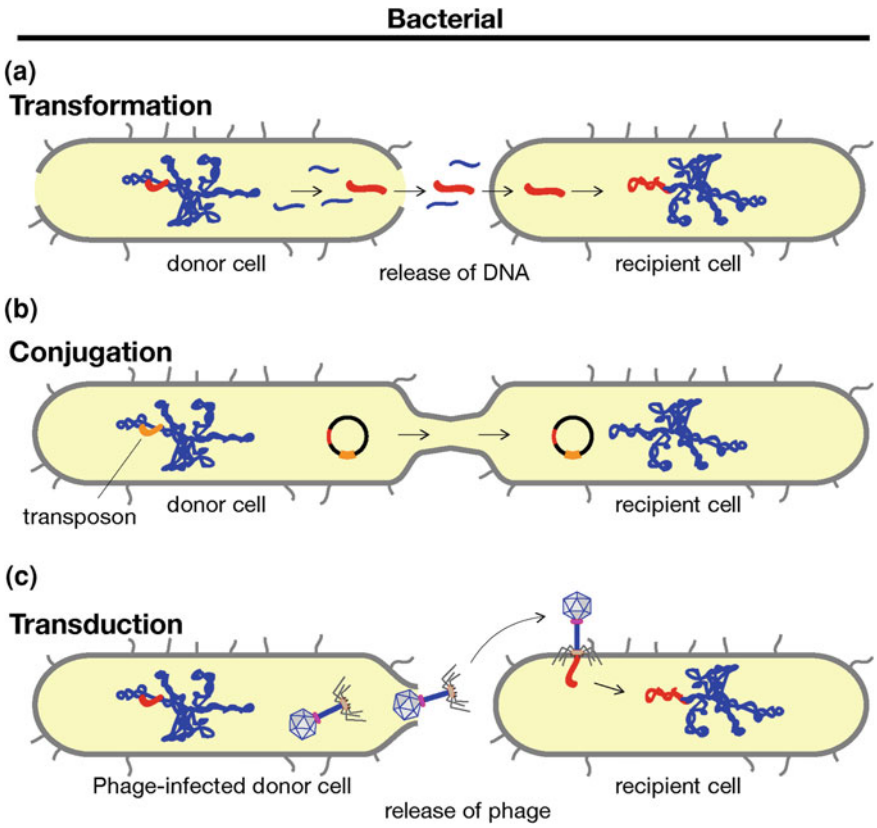


Fig. 8.2 Mechanisms of horizontal gene transfer in bacteria. **a** Uptake of naked DNA from the environment (transformation), **b** Transfer of plasmid genetic material through the mating-pore pilus from a donor to a recipient bacterial cell (conjugation), **c** Transfer of genetic material from a donor to a recipient bacterial cell through a bacteriophage intermediate (transduction)

transformation there is no actual need for the donor to be physically present either in terms of time or in terms of space. The recognition and uptake of naked DNA directly from the environment (transformation) is a widespread DNA transfer mechanism, present in many archaeal and bacterial species including Gram positive and Gram negative representatives (Lorenz and Wackernagel 1994). In order for natural transformation to occur, a physiological state of competence must be reached; some bacteria species develop competence as a response to certain environmental changes whereas others, such as *Neisseria gonorrhoeae* and *Haemophilus influenzae* are constantly competent to accept naked DNA (Dubnau 1999). Transformation in Neisseria and *H. influenzae* is selective and requires the presence of specific DNA Uptake Sequences (DUS) of approximately 10 bp in length (Goodman and Scocca 1988) that are scattered throughout the bacterial chromosome at frequencies up to 2,000 copies per chromosome (Parkhill et al. 2000).

DNA transfer between bacterial genomes can occur also through a different mechanism (i.e., transduction) that presupposes the presence of intermediates that fail to fit within the actual definition of a living organism, namely bacteriophages. Bacteriophages are viruses specialized to infect bacteria and a recent estimate suggests that approximately 10^{30} tailed bacteriophages exist on our planet, a number that far exceeds the population of any "living" organism (Brussow and Hendrix 2002). There are two major types of transduction, generalized, and specialized. In the first case, random fragments of the host bacterial chromosome can be packaged within the phage capsid during the replication and maturation process of the particles of a lytic bacteriophage. Some phage particles carry exclusively bacterial DNA, and upon a second infection they can transfer genetic material from one bacterium to another. Alternatively temperate bacteriophages integrate their genetic material into the bacterial chromosome, forming prophage elements. Upon induction a small part of the bacterial chromosome, close to the attachment site of the bacteriophage, is picked up and substitutes a small part of the actual prophage DNA; during the phage replication process the bacterial fragment replicates along with the phage DNA, such that every phage particle at the end will contain the same bacterial DNA fragment (specialized transduction). Upon a second infection, the DNA fragment of the previous host can now be transferred to a new bacterial recipient. The amount of transferable DNA through transduction depends on the actual dimension of the phage capsid and can be up to 100 kb (Ochman et al. 2000).

Different bacteriophages infect certain bacterial species, and their specificity depends on the presence of distinct cell surface receptors on the bacterial cell. The impact and extent of transduction as a mechanism of HGT can be concluded from a previous study (Canchaya et al. 2003) focused on 56 sequenced Gram-positive and Gram-negative bacteria: 71 % of those bacterial chromosomes contain at least one prophage sequence while prophages may account for up to 16 % of the bacterial chromosomal DNA (Ohnishi et al. 2001).

Conjugation is another mechanism of cell-to-cell DNA transfer that presupposes the physical co-occurrence of both the donor and the recipient cell. Conjugation is a widespread mechanism that allows the exchange of genetic material between distantly related lineages and even between different domains of life, e.g., bacteria-plant transfer (Buchanan-Wollaston et al. 1987). Conjugation frequently involves the transfer of a mobilizable or self-transmissible plasmid through a cell-to-cell bridge (mating pillus) from a donor to a recipient cell under a rolling-circle replication process (Khan 1997).

Some plasmids of Gram-negative bacteria build the mating pillus utilizing a type IV secretion system (T4SS) and the specificity of the actual conjugation is determined by several factors including the interaction of the pillus with the outer membrane and the cell surface structure of the recipient cell (Anthony et al. 1994). If prior to the conjugation event, the plasmid had been inserted within the actual chromosome of the donor, e.g., via a recombination event between sequences of the plasmid and the chromosome, it is possible for DNA fragments of the donor chromosome to be captured by the plasmid and get transferred to the recipient cell; a

subsequent recombination between the donor DNA fragment and the recipient chromosome represents the final step in the HGT event via a conjugation mechanism.

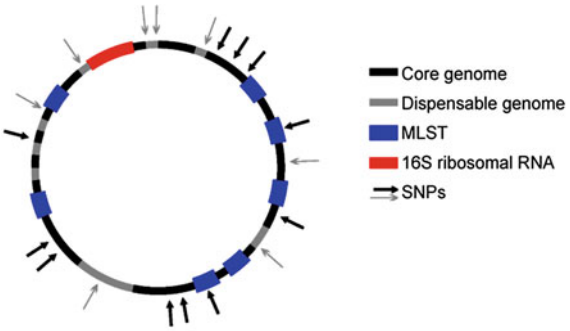
The widespread impact of HGT in blurring the boundaries of biological species has profoundly challenged the phylogenetic resolution of traditional classification systems in modern evolutionary and comparative genomics biology, calling for new, more realistic and adaptable methodologies to be exploited.

8.2 Classification Systems

In terms of phylogenetic resolution, traditional classification systems geared toward analyzing a handful of genetically distinct, often nonoverlapping species representatives are capturing only a tiny fraction (Fig. 8.3) of the species variation (Medini et al. 2008); as such they struggle to cope with the increasingly complex structure, the overlapping (fuzzy) boundaries and the dynamic nature of bacterial populations. Moving from single-gene (e.g., 16s rRNA (Woese 1987)) phylogenies trying to capture the phylogenetic history of an entire bacterial species exploiting only a tiny sequence sample ($\sim$0.07 %) of a genome, to approaches using a much larger sequence sample ($\sim$0.2 %) (e.g., multilocus sequence typing—MLST (Maiden et al. 1998)) and recently to whole-genome (Medini et al. 2005; Tettelin et al. 2005) comparative genomics (100 % coverage), is definitely a big step closer to understanding and more reliably reconstructing the phylogenetic history of bacterial populations.

The current recognition of increased microbial genome fluidity indicates that the fundamental definition of a biological species (Mayr 1942) fails in some cases to provide a realistic description of the dynamic relationships that shape microbial evolution. These findings do not support the strictly bifurcating tree of life as a means of phylogenetic analysis and instead favor the more realistic model of a phylogenetic network (Huson and Bryant 2006), which better represents the true relationships among species that are characterized by high rates of DNA exchange (Doolittle 1999a, b; Gogarten and Townsend 2005; Kunin et al. 2005).

The first data to support a reticulate (multi-furcating) model came from the genomic analysis of the obligate intracellular bacterium *Wolbachia pipientis*. Klasson et al. (2009) compared 450 genes shared by three *W. pipientis* strains (wRi, wMel, and wUni) that infect *Drosophila simulans*, *D. melanogaster*, and *Muscidifurax uniraptor*, respectively. Approximately 30 % of core genes indicated that wMel and wRi are sister lineages, a different $\sim$30 % supported the wMel and wUni sister phylogeny and 20 % showed that wRi and wUni are the more closely related pair. The authors concluded that the high rates of intra-species recombination in *W. pipientis* do not allow drawing a one-to-one relationship between gene history, genome history, and strain phenotype. This suggests that *W. pipientis* is a mixture of subpopulations, and strains in the same subpopulation recombine more frequently, which each other than with strains outside of it.



Method	Genome coverage (%)	Core genes	Dispensable genes
16S rRNA	0.07[a]	Yes	No
MLST	0.2[a]	Yes	No
SNPs	2[b]	Yes	Yes
Whole-genome	100	Yes	Yes

Fig. 8.3 Microbial genome typing systems. Properties of four methods for the comparative analysis of microbial genomes. Estimates have been calculated based on: **a** *Neisseria meningitidis*: genome size ~2.2 Mb (Bentley et al. 2007), 16S rRNA length ~1.5 kb (Sacchi et al. 2002), length of MLST loci ~4 kb (Maiden et al. 1998). **b** *Salmonella typhi*: genome size ~4.8 Mb (Deng et al. 2003), SNPs on gene fragments covering ~89 Kb (Roumagnac et al. 2006). *Source* (Medini et al. 2008)

In a second example, Didelot et al. (2007) compared the genomes of eight serovars of *Salmonella enterica* to identify blocks of high or low similarity. Their data showed that in all but one pairwise comparison the distribution of sequence divergence is unimodal. However, in the case of Paratyphi A and Typhi, the distribution showed two peaks corresponding to regions of high (1.2 %) and low (0.18 %) sequence divergence. Overall, in 75 % of their DNA sequences the two serovars appeared to be distantly related isolates of *S. enterica* and in 25 % they resemble sister lineages. The authors suggest that this apparent relatedness is the result of more than 100 recombination events that took place over a recent, short-time window.

A similar pattern of genome mosaicism is seen in *Pseudomonas fluorescens*. Silby et al. (2009) sequenced the genomes of two *P. fluorescens* strains (SBW25 and Pf0-1) and compared them with that of *P. fluorescens* Pf-5. The comparison yielded a shared core set of ~3,600 protein-coding genes, which corresponds to only ~60 % of genes in each of the three genomes. By contrast, a similar analysis of five isolates of *Pseudomonas aeruginosa* gave a core set of almost 5,000 genes, with only 1–8 % of protein-coding genes being strain specific. Despite this diversity, a comparison of the three *P. fluorescens* strains and *P. aeruginosa* PA01 showed that almost 24 % and 35 % of the genes place SBW25 closest to Pf-5 and Pf0-1, respectively, and 37 % put Pf0-1 in the same node as Pf-5, suggesting that there has been extensive genetic recombination between these strains despite their extreme diversity.

These three examples show that, in the case of highly mosaic genomes, traditional models for analyzing the history of microorganisms are not directly

applicable. Methodologies that tailor the model to the data, rather than the data to the model, offer a more realistic description of microbial diversity, dynamics and complexity (Vernikos 2009).

The obfuscating impact of HGT within species creating reticulate pathways of genetic exchange, paved the way toward the realization that the genome of a species might be much larger than the single-isolate genome sequence, leading to the introduction of a new term in comparative genomics, namely "pan-genome".

8.3 Pan-Genome

The extent of intra-species diversity in bacterial populations was underlined vividly by a study that focused on whole-genome-sequence comparisons of eight *Streptococcus agalactiae* isolates (Tettelin et al. 2005): the results revealed that the pan-genome (Medini et al. 2005)—the genome of a whole bacterial species that consists of core genes and dispensable genes that are partly shared—might be much larger than the genome of a single isolate.

Group B Streptococcus (GBS or *Streptococcus agalactiae*) pan-genome is predicted to grow by an average of 33 new genes every time a new strain is sequenced with an estimated core and dispensable gene dataset of 1,806 and 907 genes respectively. Analysis on five *Streptococcus pyogenes* predicts an asymptotic value of 27 specific genes for each new genome added. On the other side of the spectrum sits *Bacillus anthracis*, where isolates converge to zero after the addition of only a 4th genome. Hence, the *B. anthracis* species has a 'closed' pan genome, and four genome sequences are sufficient to completely characterize this species. Streptococci, Meningococci, *H. pylori*, Salmonellae and *E. coli* are likely to have an open pan-genome. On the other hand *B.anthracis*, *Mycobacterium tuberculosis* and *Chlamydia trachomatis* live in isolated niches with limited access to the global microbial gene pool and are likely to have closed pan-genome (Medini et al. 2005).

Comparison of 17 *E. coli* genomes resulted in identification of $\sim$2,200 genes conserved in all isolates. Calculations indicate that *E. coli* genomic diversity represents an open pan-genome containing a reservoir of more than 13,000 genes. The open pan-genome of *E. coli* indicates that every new genome will contribute on average 300 novel genes in the pan-genome (Rasko et al. 2008).

N. meningitidis pan-genome grows slowly because strain-specific genes are rare. The asymptotic core genome size was estimated to be 1,630 $\pm$ 62 genes. The pan-genome was confirmed as open but growing at a slow rate. Extrapolation of the data indicates that, if 100 genomes were sequenced, the *N. meningitidis* pan-genome would consist of $\sim$2,500 genes and each single isolate thereafter would contribute an average of less than two new genes. Each meningococcal genome is expected to be composed, on average, of 79 % core, 21 % dispensable, and $<$0.1 % specific genes (Budroni et al. 2011).

Despite the extreme rates of genetic exchange that bacteria frequently indulge in, and their rather fuzzy phylogenetic boundaries leading to open pan-genomes, their self-integrity is preserved over-time via dedicated, highly selective, and dynamic "machineries" such as restriction modification systems.

8.4 Restriction Modification Systems

Restriction-modification (RM) systems (Wilson and Murray 1991; Bickle and Kruger 1993; Heitman 1993; Raleigh and Brooks 1998) consist of two active components: a methyltransferase (Cheng 1995; Jeltsch 2002) modifying adenine or cytosine residues at specific recognition sites and a restriction endonuclease (Pingoud and Jeltsch 1997, 2001) that recognizes the same sequence pattern and slices the DNA if it is in an unmethylated state. More than 200 different systems have been identified so far in bacteria (Roberts and Macelis 2001), while a large number of different RM-systems occur in single species (e.g. 16 different RM-systems in *Neisseria gonorrhohae* (Stein et al. 1995)). In terms of specificity, almost all palindromic sequences comprising 4 or 6 bp can be recognized by at least one RM-system.

One of the first potential roles assigned to RM-systems was the defense mechanism against bacteriophage attack. From the phage point of view, there are several strategies to avoid the host's RM systems e.g. by reducing the number of RM-specific DNA patterns (Bickle and Kruger 1993). Similar strategies are exploited during plasmid conjugation, e.g., via anti-restriction proteins acting as RM-inhibitors (Velkov 1999).

More recently, an additional pivotal role of RM-systems has started to emerge: RM-systems can act as genetic flux "switches" securing the maintenance of phylogenetic self-integrity. Living on the fast lane of evolution and exchanging genetic material with other taxa (via HGT), bacteria run the risk of de-speciation or phylogenetic obfuscation. Preserving the genetic integrity of a species requires a fine-tuning, highly selective process securing some level of genetic isolation, which in return acts a driving force for the evolution of species. Genetic isolation can be achieved via different strategies and processes like geographic isolation in higher organisms; in bacteria that indiscreetly exchange genetic material, controlling the very process of DNA uptake from the environment is a possible way of establishing selective genetic isolation (Tortosa and Dubnau 1999).

Distinct and diverge DNA methylation patterns that occur in bacterial genomes enable the recognition of native and "alien" DNA. Such a barcoding recognition process has already been assigned to RM-systems and endonucleases that cleave methylated DNA as a mean to control the genetic flux involving DNA originating from other species or taxa (Murray et al. 1975). The level of genetic isolation can be extremely fine-tuned enabling high intra- and/or inter-species resolution, by hosting many different RM-systems within single bacterial species. A good example is *N. meningitidis*, which consists of two biotypes, one containing a dam methyltransferase which generates methylated GmATC sites and another

containing the *drg* restriction enzyme that cleaves these sites (Bucci et al. 1999). RM-driven isolation from the genetic background of the population can cause evolution in quantum leaps leading either into the emergence of new species or the extinction of another. This selective mechanism is so tightly connected and streamlined to the survival strategy of each bacterium that in situation of hostile environmental conditions where there is "desperate" need of "alien" DNA and genome "rejuvenation", RM-systems turn off (Velkov 1999).

It becomes obvious that bacteria, with their vast and rather exotic armory, "walk" on the edges of biological viability, living on vertical or horizontal gene flow lifestyles exploiting practically the entire spectrum of any conceivable population structure.

8.5 Population Structure Spectrum

Bacterial population structure dynamics range from clonal (e.g., *E. coli*, with occasional background HGT), to panmictic where rates of HGT are so high that genetic relationships between taxa are shuffled to such an extent that pure phylogenetic traces become invisible (e.g., *N. gonorrhoeae*). Within this wide and extreme spectrum, there are a few bugs that indulge both worlds, i.e., clonality and genome fluidity via horizontal genetic transmission. The overwhelming phylogenetic signal of recombination shapes the backbone of such epidemic populations, where distinct genotypes "mate" indiscreetly and vigorously. Occasionally, frequent genotypes, or clusters of closely related genotypes emerge and persist with high frequency in the population. Such "average" population structures have been named as "epidemic structures" in a seminal work by John Maynard Smith and co-workers (Smith et al. 1993). Multi Locus Sequence Type analysis showed that the population of *N. meningitidis* provides a typical example of epidemic structure. In such "intermediate" population structures groups of closely related genotypes, named clonal complexes, persist in time and spread geographically despite the effect of homologous recombination. Different models have been proposed to account for this apparent contradiction, including fitness landscapes (Feil 2004), neutral microepidemic evolution (Fraser et al. 2005) and immune selection (Buckee et al. 2008).

8.6 Speciation, Phylogenetic Structure and Genome Stability of *N. meningitidis*

The genome of *N. meningitidis* has both signatures of clonal descent and HGT, supporting the presence of multiple distinct genotypes in the population (Caugant et al. 1986; Smith et al. 1993). Multilocus sequence typing (MLST) resolves the population structure of *N. meningitidis* in distinct clonal complexes (CC), based on the sequence similarity of neutral loci (Maiden et al. 1998); these CCs persist in the population for many decades, despite the high rates of recombination (Feil

et al. 2001; Fraser et al. 2005; Jolley et al. 2005). Recent evidence suggests though, that this seemingly neutral pattern of evolution that could well explain the presence of distinct lineages in the population, oversimplifies the frequent and extensive patterns of intra and inter-species variation and different hypotheses have been put forward (Fraser et al. 2005; Jolley et al. 2005; Buckee et al. 2008).

Based on recent comparative genomic analysis (Bentley et al. 2007; Schoen et al. 2009), the species of *N. meningitidis* differentiated from the genus, via the acquisition of distinct insertion sequences and capsular polysaccharide genes by an unencapsulated ancestor and classification groups larger than the CCs namely phylogenetic clades (PCs) have been proposed to more reliably capture the dynamic and "2-geared" (expand and contract) evolution pattern in the population of *N. meningitidis* (Schoen et al. 2008). The aforementioned genetic flux "switches" seem to sit at the basis of these PCs, in which distinct set of restriction modification systems selectively block "alien" DNA sequence, securing a higher level population structuring than initially proposed via MLST data.

One of the driving forces of homologous recombination either toward the direction of creating cell-surface variability—a key determinant of host-pathogen interaction—or toward the direction of preserving genome stability and phylogenetic sanity is the presence of numerous diverse families of repeat arrays scattered throughout the entire genome sequence of *N. meningitidis*.

8.7 Repeat Arrays and Genome Fluidity

N. meningitidis genome contains many hundreds of repetitive sequence elements ranging from simple sequence repeats associated with phase variable genes, to complete gene cluster duplications. Variation of the bacterial cell surface is among others, driven by specific genes and associated repetitive DNA sequences. The repeat sequences promote the swapping of genes that code for variant copies of cell surface components. This dynamic cell-surface variation in return seems to dictate profoundly the host-pathogen interaction (Bentley et al. 2007).

DNA uptake sequences (involved in the uptake of naked DNA from the environment, i.e. transformation) are the most abundant ($\sim$1,900/genome) repeats and are scattered throughout the genome (Goodman and Scocca 1988). The next most frequent repeat family is that of NIMEs (Neisserial Intergenic Mosaic Elements): 20-bp inverted repeats, namely dRS3 elements, flanking over 100 families of $\sim$100-bp repeat sequences, namely RS elements; NIMEs are often clustered into long arrays of multiple dRS3s separated by different RS elements (Parkhill et al. 2000). Another frequent repeat family is that of Correia elements comprised of conserved repeat sequences ($\sim$150 bp in length) bounded by 51-bp inverted repeats. Correia elements are often located upstream of coding sequences (Liu et al. 2002) and may affect gene expression (De Gregorio et al. 2002; Packiam et al. 2006).

It has been hypothesized that NIME arrays may encourage sequence variation in neighboring genes by increasing the frequency of recombination with

exogenous DNA, via homologous or site-specific recombination (Parkhill et al. 2000). The consistent chromosomal position and the variable length of these repeat arrays in different serogroups, points toward a hypothesis of common ancestry introduction on the one hand and on the other hand toward a dynamic structure (contracting and expanding), as a result of recurring recombination.

The average % identity between orthologous genes flanking repeat arrays is significantly lower than the average % identity of orthologues not flanking repeat arrays, putting forward the hypothesis that, in addition to the immune selection preserving variants, repeat arrays boost diversity in flanking sequences via recombination with alien DNA, increasing the rate of gene exchange at the adjacent loci. In support of this hypothesis, there is a clear association between repeat arrays and genes encoding cell surface proteins where increased sequence variation could be advantageous in host interactions. The majority of flanking genes code for cell surface or exported proteins, regardless of the length of the repeat array and this proportion seem to increase relative to the array length.

8.8 *Neisseria Meningitidis* Population: PCs Hosting Different CCs

The higher level phylogenetic structuring, became evident via reticulate network analysis, whereby the strictly bifurcating pattern of evolution is not a restrictive factor, and both clonal and horizontal genetic flux can be simultaneously mapped in the model structure of the network. The network analysis revealed that indeed strains from the same CC are closely related forming monophyletic groups but additionally, strains from distinct CCs, i.e., CC32/CC269, CC8/CC11, and ST-41/ST-44 subcomplexes group together at a higher level under three PCs, respectively, namely: PC32/269, PC8/11, PC41/44, indicating common ancestry.

8.9 PC-Specific Chromosomal Rearrangements

Within 11 *N. meningitidis* genome sequences, ten major chromosomal rearrangements stand out with breakpoints mostly associated with dRS3 repeats and IS elements and sufficient to reconstruct the collinearity of the chromosomes. Three rearrangements most probably happened more than one time, seven rearrangements were predicted to have occurred only once, and six were PC-specific, further supporting the evolutionary consistency and phylogenetic robustness of PCs. Seven inversions have a potential biological impact on the chromosomal regions flanking the breakpoints, four of which might influence the expression of RM and virulence-related genes.

Half of the breakpoints are related to one or more dRS3 elements; four breakpoints are flanked by an IS, two are flanked by DUSs, two are flanked by a

complex repeated region, and two are flanked by an rRNA. Genome-wide repeat density analysis, showed significant enrichment/depletion of dRS3 and DUS elements, respectively. On average, in each genome, the dRS3 density is 3.1 ± 0.2 elements per 10 kb, but in the breakpoint regions, this density grows to an average of 5.4 ± 0.3. Conversely, a deficit of DUSs was observed in breakpoint-associated regions, and an inverse association between DUS and dRS3 elements was measured on the whole chromosome suggesting DUS replacement by dRS3s.

These rearrangements are likely to have functional impact in the chromosome, potentially affecting distinct and diverse cell functions, including DNA uptake and sequence variation via recombination, adhesion, and penetration to human host cells, colonization and invasion, induction of bactericidal antibodies, pilus biosynthesis and retraction, transformation competence, generation of antigen diversity, host-pathogen interaction, capsule expression, and bacteriocin resistance.

8.10 Host–Pathogen Interaction via Highly Conserved Clade-Specific Genes

Each PC has distinct PC-specific gene pool and in contrast with what would be expected simply by chance, these PC-specific gene tanks have more genes in common compared to smaller groups at the CC-level. In other words, more PC-specific than CC-specific genes exist and almost all 20 regions (eight regions for PC32/269, four regions for PC8/11, and eight regions for PC41/44) are extremely highly conserved (nucleotide sequence identity close to 100 %), implying a recent emergence of PCs. PC-specific gene pools include RMs genes and genes involved in host-pathogen interaction, that were derived either via HGT or local genomic rearrangements, differentiating further, at the functional level, the PCs in *N. meningitidis* population.

In PC32/269, six strains share eight PC-specific regions hosting 13, highly conserved (sequence identity >99.8 %) genes and cover a wide functionality range from a two-partner secretion (TPS) systems involved in secretion of large virulence-related proteins contributing to adhesion to epithelial cells (pronounced in invasive meningococcal CCs (Schmitt et al. 2007; van Ulsen et al. 2008)), cassettes encoding putative variants of the C-terminal ends of hemagglutinin contributing to variation via genetic recombination (Bentley et al. 2007), to zinc uptake regulator that represses ZunD, a vaccine candidate that elicits reactive antibodies in humans (Gaballa and Helmann 1998; Smith et al. 2009; Stork et al. 2010).

In PC8/11, five strains have four PC-specific chromosomal regions in common, sharing eight highly conserved genes (sequence identity >99.9), with potential functionality ranging from DNA exchange via homologous recombination, cell mobility, exogenous DNA uptake, host-pathogen interaction (Chen and Dubnau 2004; Carbonnelle et al. 2009), to survival on exposure to stress (Fivian-Hughes and Davis 2010).

The four strains in PC41/44 share eleven highly conserved (sequence identity ≥99.8 %) genes organized in eight PC-specific chromosomal regions of various lengths possibly involved in DNA exchange via homologous recombination, biosynthesis and degradation of surface polysaccharides and lipopolysaccharides, potentially conferring PC-associated capsular specificities, iron-uptake from the environment, pillus-biosynthesis, to competence (DNA receptor and binding).

8.11 Homologous Recombination Pervasiveness and DNA Uptake Sequences

Phylogenetic networks reveal homoplasy in the form of nontree-like edges, horizontal phylogenetic signal was predicted to be confined to a very limited number of DNA donor–acceptor pairs and homologous recombination was detected in 87 % of each chromosome. No significant positional bias for recombination was detected along the chromosome, and the rate of detectable recombination ρ did not correlate positively with the degree of sequence conservation; this suggests that recombination acts similarly on most of the genome. A significant correlation was found between ρ and the density of DUSs and a smaller proportion of nonrecombining DNA was predicted in the core genome (11 %) than in the dispensable genome (45 %), where DUS density is much lower (Treangen et al. 2008). These results confirm the link between DUSs and homologous recombination and the role of the latter in preserving genome stability rather than generating adaptive variation (Treangen et al. 2008).

8.12 Insertion Sequences Violate PCs Boundaries

39 IS types, belonging to nine families, were detected in variable copy numbers in each genome. On average, each genome hosts 41 ISs distributed evenly across the genomes. With a few exceptions analysis showed that ISs move quite freely within the species, frequently crossing PC borders; however IS-based clustering segregates clearly meningitidis species from the rest of the genus (Schoen et al. 2008), suggesting that IS-based phylogenetic resolution is low and discriminative only at the species level.

8.13 RMSs Shape and Preserve PCs' Self-integrity

PC-specific signatures could only be identified in RMS-related genes or positional rearrangements. 22 putative RMSs were identified (Budroni et al. 2011), including 14 Type II, 4 Type III, and 2 Type I systems. No RMS is global, i.e., present in all

strains, and on average five to nine RMSs are present in each genome. Two RMSs are found in all but one analyzed isolates. There are eight isolate-specific RMSs, and five are unique to a capsule-null strain. 2–13 isolates share the rest of 12 RMSs and one-half of the RMSs are localized in HGT integration site hotspots. Most of RMSs have a GC% deviating from the *N. meningitidis* native backbone composition, putting forward an "alien" origin of the RMSs, possibly acquired via HGT from other taxa.

The phyletic profile (i.e., gene presence or absence) of RMSs, in contrast to IS elements, reliably (bootstrap values: 92–100 %) reconstructs the species genomic phylogeny; the three PCs (PC32/269, PC41/44 and PC8/11) host a unique combination of seven, nine and seven RMSs, respectively. It is worth noting that it is the very specific combination of RMSs that differentiates uniquely the three clades and not each RMS individually. An extensive (189 strains) meta-analysis confirmed the validity of these findings.

8.14 HGT Flows in Larger Quanta Intra-PC Compared to Inter-PC

Analysis of between and within PCs gene transfer (Budroni et al. 2011), differs significantly in terms of DNA quantity (i.e., bp length) and not in terms of number of events, based on a 20 genome collection. The average length of intra and inter-PC gene flow is 3.89 and 0.68 kb, respectively. This observation further supports the pivotal role of RMSs in driving and preserving the population structure of *N. meningitidis* since differentiation at the DNA quantity and not at the number of HGT events, can be explained by a highly selective genetic flux "switch" acting at the post-uptake and pre-integration step of "alien" DNA arrivals.

8.15 Discussion and Outlook

HGT is a pivotal mechanism of microbial evolution. In several bacterial species the clonal mutation patterns typical of asexual reproduction are obfuscated by recombination that generates panmictic populations. In some species, however, cohesive groups of genotypes named clonal complexes (CCs) persist in space and time despite high rates of HGT.

N. meningitidis, a pathogenic bacterium prototypical of such "intermediate" patterns, was shown to be structured in phylogenetic units larger than CCs, named Phylogenetic Clades (PCs). Using CCs to answer a common ancestry question, suffers from low phylogenetic resolution (limited number/size of loci) to be able to capture a reliable and representative species signal, while whole-genome phylogenies (evaluated via reticulate networks) proved more promising in drawing cross CC relationships.

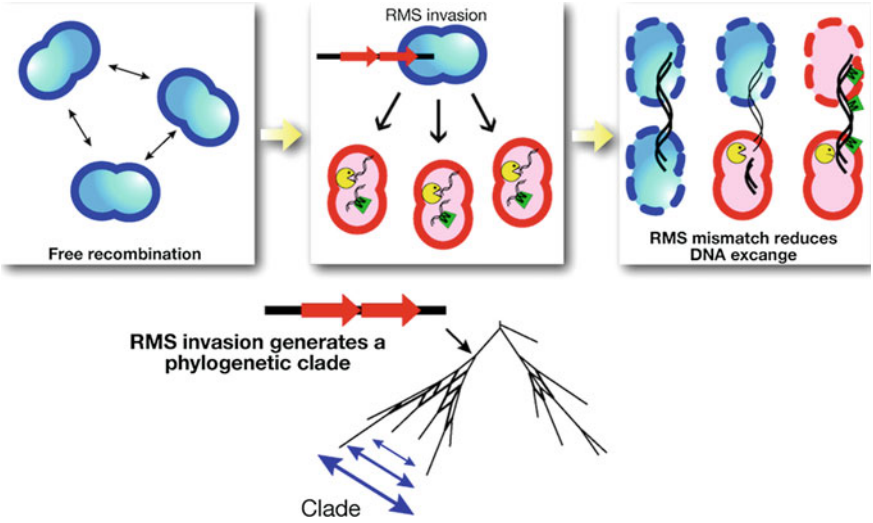


Fig. 8.4 Horizontal gene transfer modulated by restriction-modification systems as the root cause for population structuring in bacteria. Working model for RMS-driven origin and persistence of Phylogenetic Clades in *Neisseria meningitidis*. *Source* (Budroni et al. 2011)

Three PCs identified, host specific gene content and arrangements driving potentially host-pathogen interactions. *N. meningitidis* distinct population structuring seems to contradict the high rate of gene flow, and a putative highly selective genetic flow "switch" role has been assigned to the 22 RMSs identified in 20 strains, to explain this contradicting dynamic interplay; worth noting is the fact that the phyletic profile of these systems, seem to reliably reconstruct the species phylogeny. The pivotal role of abundant repeat families (mostly DUSs) to drive high rates of homologous recombination further supports the average of 1.6 recombination events per mutation event.

So how can the RMS and PCs phylogenetic "alignment" be best explained? Is this a driving force of phylogenetic structuring or the consequence of diversifying evolution? Simply by chance, one would expect a higher rate of recombination among closely related isolates, due to sequence similarity, compared to more distantly related ones. Counterintuitively though, this is not the case for PCs and CCs where there are more PC-specific than CC-specific genes, although the former are much larger phylogenetic groups than the latter; in other words, there is no correlation between rate of recombination and sequence similarity of isolates. Moreover, gene flow events occur five times more frequently intra than inter-PC and the size of DNA exchange correlates with the number of RMSs in common between donor and acceptor strains. These data put forward a possible PC-specific DNA cleavage, RMS-driven mechanism whereby "alien" DNA is cleaved after its arrival in the acceptor cell and prior to its integration in the new chromosome. Similar mechanisms have been proposed in *Helicobacter pylori* (Lin et al. 2009) and *Haemophilus influenza* (Erwin et al. 2008).

It turns out that in *N. meningitidis* population, counterintuitively homologous recombination, instead of obfuscating the population structure, is the very cause of the PCs structuring. As shown in Fig. 8.4, in a panmictic background, whereby phylogenetic signature weakens due to homologous recombination, new clones can emerge by chance via HGT acquisition of RMSs. This could well form the very first step, toward "adolescence" and differentiation and via offspring inheritance of this new DNA "legacy", the progenitor of the clone is less affected by the homogenizing effect of homologous recombination in the background population, while its offspring indulges more eagerly into "closed-door" genetic exchange, giving over time rise to a new distinct lineage in the population.

Acknowledgments We thank Giorgio Corsi for artwork and figure preparation.

References

Anderson NG (1970) Evolutionary significance of virus infection. Nature 227:1346–1347

Andersson JO, Doolittle WF, Nesbo CL (2001) Genomics. Are there bugs in our genome? Science 292:1848–1850

Anthony KG, Sherburne C, Sherburne R, Frost LS (1994) The role of the pilus in recipient cell recognition during bacterial conjugation mediated by F-like plasmids. Mol Microbiol 13:939–953

Baumler AJ (1997) The record of horizontal gene transfer in Salmonella. Trends Microbiol 5:318–322

Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T et al (2007) Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet 3:e23

Bickle TA, Kruger DH (1993) Biology of DNA restriction. Microbiol Rev 57:434–450

Bohr N (1913) On the constitution of atoms and molecules. Phil Mag 26:1–15

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL et al (2003) Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328

Brochier C, Philippe H, Moreira D (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet 16:529–533

Brussow H, Hendrix RW (2002) Phage genomics: small is beautiful. Cell 108:13–16

Bucci C, Lavitola A, Salvatore P, Del Giudice L, Massardo DR, Bruni CB et al (1999) Hypermutation in pathogenic bacteria: frequent phase variation in meningococci is a phenotypic trait of a specialized mutator biotype. Mol Cell 3:435–445

Buchanan-Wollaston V, Passiatore JE, Cannon F (1987) The mob and oriT mobilization functions of a bacterial plasmid promote its transfer to plants. Nature 328:172–175

Buckee CO, Jolley KA, Recker M, Penman B, Kriz P, Gupta S et al (2008) Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. Proc Natl Acad Sci USA 105:15082–15087

Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C et al (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci USA 108:4494–4499

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H (2003) Phage as agents of lateral gene transfer. Curr Opin Microbiol 6:417–424

Carbonnelle E, Hill DJ, Morand P, Griffiths NJ, Bourdoulous S, Murillo I et al (2009) Meningococcal interactions with the host. Vaccine 27(Suppl 2):B78–B89

Caugant DA, Froholm LO, Bovre K, Holten E, Frasch CE, Mocca LF et al (1986) Intercontinental spread of a genetically distinctive complex of clones of *Neisseria meningitidis* causing epidemic disease. Proc Natl Acad Sci USA 83:4927–4931

Chen I, Dubnau D (2004) DNA uptake during bacterial transformation. Nat Rev Microbiol 2:241–249

Cheng X (1995) Structure and function of DNA methyltransferases. Annu Rev Biophys Biomol Struct 24:293–318

Choi J, Shin D, Ryu S (2007) Implication of quorum sensing in *Salmonella enterica* serovar typhimurium virulence: the luxS gene is necessary for expression of genes in pathogenicity island 1. Infect Immun 75:4885–4890

Darwin C (1859) On the origin of species by means of natural selection. J. Murray, London

Daubin V, Moran NA, Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. Science 301:829–832

De Gregorio E, Abrescia C, Carlomagno MS, Di Nocera PP (2002) The abundant class of nemis repeats provides RNA substrates for ribonuclease III in Neisseriae. Biochim Biophys Acta 1576:39–44

Deng W, Liou SR, Plunkett G 3rd, Mayhew GF, Rose DJ, Burland V et al (2003) Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J Bacteriol 185:2330–2337

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R et al (2002) The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol 4:453–461

Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D (2007) A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res 17:61–68

Doolittle RF, Feng DF, Anderson KL, Alberro MR (1990) A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J Mol Evol 31:383–388

Doolittle RF, Feng DF, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Doolittle WF (1999a) Lateral genomics. Trends Cell Biol 9:M5–M8

Doolittle WF (1999b) Phylogenetic classification and the universal tree. Science 284:2124–2129

Dubnau D (1999) DNA uptake in bacteria. Annu Rev Microbiol 53:217–244

Dunning Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P, Munoz Torres MC et al (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756

Erwin AL, Sandstedt SA, Bonthuis PJ, Geelhood JL, Nelson KL, Unrath WC et al (2008) Analysis of genetic relatedness of *Haemophilus influenzae* isolates by multilocus sequence typing. J Bacteriol 190:1473–1483

Feil EJ (2004) Small change: keeping pace with microevolution. Nat Rev Microbiol 2:483–495

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC et al (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci USA 98:182–187

Fivian-Hughes AS, Davis EO (2010) Analyzing the regulatory role of the HigA antitoxin within *Mycobacterium tuberculosis*. J Bacteriol 192:4348–4356

Fraser C, Hanage WP, Spratt BG (2005) Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci USA 102:1968–1973

Gaballa A, Helmann JD (1998) Identification of a zinc-specific metalloregulatory protein, Zur, controlling zinc transport operons in *Bacillus subtilis*. J Bacteriol 180:5815–5821

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ et al (2005) Opinion: re-evaluating prokaryotic species. Nat Rev Microbiol 3:733–739

Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679–687

Goldman BS, Kranz RG (1998) Evolution and horizontal transfer of an entire biosynthetic pathway for cytochrome c biogenesis: Helicobacter, Deinococcus, Archae and more. Mol Microbiol 27:871–873

Goodman SD, Scocca JJ (1988) Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. Proc Natl Acad Sci USA 85:6982–6986

Gribaldo S, Lumia V, Creti R, Conway de Macario E, Sanangelantoni A, Cammarano P (1999) Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 181:434–443

Griffith F (1928) The significance of Pneumococcal types. J Hyg (Lond) 27:113–159

Groisman EA, Ochman H (1996) Pathogenicity islands: bacterial evolution in quantum leaps. Cell 87:791–794

Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23:1089–1097

Heitman J (1993) On the origins, structures and functions of restriction-modification enzymes. Genet Eng (NY) 15:57–108

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Jeltsch A (2002) Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. ChemBioChem 3:274–293

Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MC (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol Biol Evol 22:562–569

Khan SA (1997) Rolling-circle replication of bacterial plasmids. Microbiol Mol Biol Rev 61:442–455

Klasson L, Westberg J, Sapountzis P, Naslund K, Lutnaes Y, Darby AC et al (2009) The mosaic genome structure of the Wolbachia wRi strain infecting *Drosophila simulans*. Proc Natl Acad Sci USA 106:5725–5730

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: reconstructing the microbial phylogenetic network. Genome Res 15:954–959

Kurland CG (2000) Something for everyone. Horizontal gene transfer in evolution. EMBO Rep 1:92–95

Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci USA 100:9658–9662

Lawrence JG (2002) Gene transfer in bacteria: speciation without species? Theor Popul Biol 61:449–460

Lawrence JG, Hendrickson H (2003) Lateral gene transfer: when will adolescence end? Mol Microbiol 50:739–749

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397

Lederberg J (1956) Conjugal pairing in *Escherichia coli*. J Bacteriol 71:497–498

Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:e130

Lin EA, Zhang XS, Levine SM, Gill SR, Falush D, Blaser MJ (2009) Natural transformation of helicobacter pylori involves the integration of short DNA fragments interrupted by gaps of variable size. PLoS Pathog 5:e1000337

Linehan SA, Rytkonen A, Yu XJ, Liu M, Holden DW (2005) SlyA regulates function of Salmonella pathogenicity island 2 (SPI-2) and expression of SPI-2-associated genes. Infect Immun 73:4354–4362

Liu SV, Saunders NJ, Jeffries A, Rest RF (2002) Genome analysis and strain comparison of correia repeats and correia repeat-enclosed elements in pathogenic Neisseria. J Bacteriol 184:6163–6173

Lorenz MG, Wackernagel W (1994) Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev 58:563–602

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R et al (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145

Mayr E (1942) Systematics and the origin of species, vol 334. Columbia University Press, New York, p 324

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R et al (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6:419–430

Morse ML, Lederberg EM, Lederberg J (1956) Transduction in *Escherichia Coli* K-12. Genetics 41:142–156

Murray K, Murray NE, Brammar WJ (1975) Restriction enzymes and the cloning of eukaryotic DNA. FEBS 10:193–207

Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ et al (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science 313:236–238

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH et al (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329

Nomura M (1999) Engineering of bacterial ribosomes: replacement of all seven *Escherichia coli* rRNA operons by a single plasmid-encoded operon. Proc Natl Acad Sci USA 96:1820–1822

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Ochman H, Wilson AC (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86

Ohnishi M, Kurokawa K, Hayashi T (2001) Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? Trends Microbiol 9:481–485

Packiam M, Shell DM, Liu SV, Liu YB, McGee DJ, Srivastava R et al (2006) Differential expression and transcriptional analysis of the alpha-2,3-sialyltransferase gene in pathogenic Neisseria spp. Infect Immun 74:2637–2650

Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR et al (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404:502–506

Pingoud A, Jeltsch A (1997) Recognition and cleavage of DNA by type-II restriction endonucleases. Eur J Biochem 246:1–22

Pingoud A, Jeltsch A (2001) Structure and function of type II restriction endonucleases. Nucleic Acids Res 29:3705–3727

Raleigh EA, Brooks JE (1998) Restriction modification systems: where they are and what they do. In: De Bruijn FJ, Lupski JR, Weinstock GM (ed) Bacterial genomes. Chapman & Hall, New York, pp 78–92

Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P et al (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 190:6881–6893

Roberts RJ, Macelis D (2001) REBASE–restriction enzymes and methylases. Nucleic Acids Res 29:268–269

Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT et al (2006) Evolutionary history of *Salmonella typhi*. Science 314:1301–1304

Sacchi CT, Whitney AM, Reeves MW, Mayer LW, Popovic T (2002) Sequence diversity of *Neisseria meningitidis* 16S rRNA genes and use of 16S rRNA gene sequencing as a molecular subtyping tool. J Clin Microbiol 40:4520–4527

Schmitt C, Turner D, Boesl M, Abele M, Frosch M, Kurzai O (2007) A functional two-partner secretion system contributes to adhesion of *Neisseria meningitidis* to epithelial cells. J Bacteriol 189:7968–7976

Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T et al (2008) Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. Proc Natl Acad Sci USA 105:3473–3478

Schoen C, Tettelin H, Parkhill J, Frosch M (2009) Genome flexibility in *Neisseria meningitidis*. Vaccine 27(Suppl 2):B103–B111

Schrödinger E (1926) An undulatory theory of the mechanics of atoms and molecules. Phys Rev 28:1049–1070

Silby MW, Cerdeno-Tarraga AM, Vernikos GS, Giddens SR, Jackson RW, Preston GM et al (2009) Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. Genome Biol 10:R51

Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–4388

Smith KF, Bibb LA, Schmitt MP, Oram DM (2009) Regulation and activity of a zinc uptake regulator, Zur, in *Corynebacterium diphtheriae*. J Bacteriol 191:1595–1603

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411:940–944

Stein DC, Gunn JS, Radlinska M, Piekarowicz A (1995) Restriction and modification systems of *Neisseria gonorrhoeae*. Gene 157:19–22

Stork M, Bos MP, Jongerius I, de Kok N, Schilders I, Weynants VE et al (2010) An outer membrane receptor of *Neisseria meningitidis* involved in zinc acquisition with vaccine potential. PLoS Pathog 6:e1000969

Syvanen M (1985) Cross-species gene transfer; implications for a new theory of evolution. J Theor Biol 112:333–343

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ et al (2002) 50 million years of genomic stasis in endosymbiotic bacteria. Science 296:2376–2379

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL et al (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955

Tortosa P, Dubnau D (1999) Competence for transformation: a matter of taste. Curr Opin Microbiol 2:588–592

Treangen TJ, Ambur OH, Tonjum T, Rocha EP (2008) The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biol 9:R60

van Ulsen P, Rutten L, Feller M, Tommassen J, van der Ende A (2008) Two-partner secretion systems of *Neisseria meningitidis* associated with invasive clonal complexes. Infect Immun 76:4649–4658

Velkov VV (1999) How environmental factors regulate mutagenesis and gene transfer in microorganisms. J Biosci 24:529–559

Vernikos GS (2009) Of trees and networks. Nat Rev Microbiol 7:691

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D et al (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99:17020–17024

Wilson GG, Murray NE (1991) Restriction and modification systems. Annu Rev Genet 25:585–627

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396

Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 9:689–710

Wyborn NR, Stapleton MR, Norte VA, Roberts RE, Grafton J, Green J (2004) Regulation of *Escherichia coli* hemolysin E expression by H-NS and Salmonella SlyA. J Bacteriol 186:1620–1628

Yap WH, Zhang Z, Wang Y (1999) Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol 181:5201–5209

Chapter 9
Quartet Partitioning Reveals Hybrid Origins of the Vertebrate

Michael Syvanen, Bryan Ericksen, Simone Linz and Jonathan Ducore

Abstract It is generally accepted that humans and sea urchins are deuterostomes and that fruit flies and jelly fish are outgroups. However, when we analyzed proteins from the genomes of these four species and submitted them to 4 taxa phylogenetic analysis, we found that, while as expected, most of the proteins (563) supported the notion of human and sea urchin in one clade and jelly fish and fruit flies in the other clade (Tree1), a large number of proteins (353) showed human and fruit fly in one clade with the sea urchin and jelly fish in the other (Tree3). Homologs were found in the genomes from 5 other metazoa. Tree1 proteins resulted in the expected 9 taxa tree, while the Tree3 proteins show vertebrates, to the exclusion of the other chordates, in the protostome clade. The two 9 taxa trees were fused into a single most parsimonious net that supports an introgression event between a vertebrate ancestor and a primitive protostome.

M. Syvanen (✉) · B. Ericksen
Department of Microbiology, University of California at Davis School of Medicine, Davis, CA 95617, USA
e-mail: syvanen@ucdavis.edu

B. Ericksen
e-mail: ericksen.b@gmail.com

S. Linz
Department of Computer Science, Center for Bioinformatics (ZBIT), University of Tübingen, Sand 14, 72076 Tübingen, Germany
e-mail: simone_linz@yahoo.de

J. Ducore
Department of Pediatrics, University of California at Davis School of Medicine, Davis, CA 95617, USA
e-mail: jmducore@ucdavis.edu

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_9,

9.1 Introduction

Fossil metazoan phyla appeared over a relatively short period of time 540 million years ago (MYA), an event called the Cambrian explosion or the metazoan radiation. Figure 9.1 shows that major metazoan assemblages appeared in the fossil record 540 MYA. If modern metazoan phyla radiated from a single point, then it would not be possible to assemble the various phyla in higher taxonomic assemblages. However, with the advent of phylogenomics which enabled hundreds if not thousands of genes to be analyzed, it became clear that the radiation occurred over a longer period of time than had been appreciated and that diversification began well before any recognizable metazoans could be seen in fossil record (Wray et al. 1996; Douzery et al. 2004; Blair and Hedges 2005; Philippe et al. 2009; Osigus et al. 2013) as is shown in Fig. 9.1. Thus, it is a realistic goal to reconstruct the pre-Cambrian evolutionary relationships of those taxa that gave rise to the modern metazoa.

In 1985, one of us offered the conjecture that horizontal gene transfer events were a major factor during the emergence of the metazoan phyla as indicated by the widespread occurrence of parallelism in the fossil record (Syvanen 1985).

Since then, a number of horizontal gene transfer events have been documented in metazoans as reviewed by Syvanen (2012). Most of the published examples involve transfers from bacteria or fungi into animals. Documenting gene transfer between metazoan phyla, a type of transfer required to explain parallel evolution implicit in Fig. 9.1, is a much more difficult problem, especially if the transfer events occurred deep in time. Earlier, we presented evidence for a possible major gene transfer event. It was found that the genome of the tunicate *Ciona intestinalis* (Sea squirt) consists of two sets of genes that support two different phylogenies (Syvanen and Ducore 2010). The simplest explanation for this result is that *C. intestinalis* descended from a hybrid, one donor being a chordate and the other belonging to an extinct phyla – likely a sister to primitive protostomes. Quartet partitioning was used to identify the proteins that fell into one or the other of these two groups; we use this method and extend it in the current study. Quartet partitioning has found application in analyzing reticulate evolution (Huson and Bryant 2006; Gauthier and Lapointe 2007), most prominently in identifying relatively recent introgressions between plant species.

9.2 Result

9.2.1 Quartet Partitions

The number of variables required to test alternative trees can be minimized by analyzing four taxa since there are only three competing unrooted trees and a single internal branch (Gaut and Lewis 1995; Hillis and Huelsenbeck 1992;

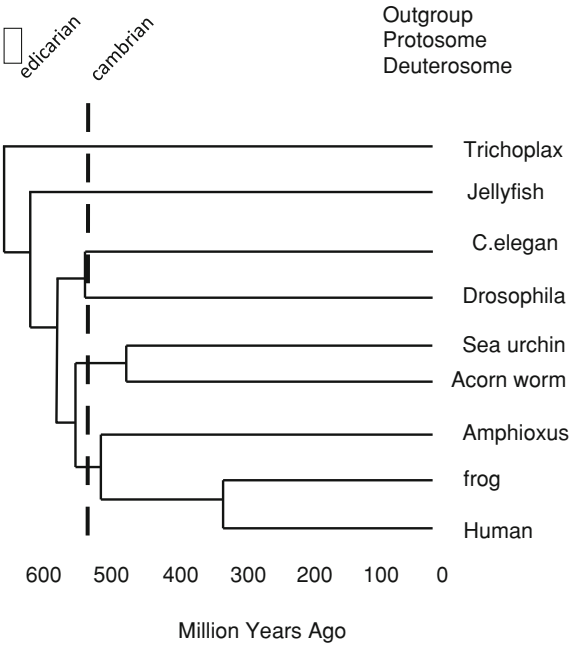


Fig. 9.1 Generally recognized relationships among metazoan phyla. The chronology is based on the geological record and the displayed taxa are the ones used in the current study

Syvanen 2002). We have chosen the jelly fish, *Nematostella vectensis*, *H. sapiens*, the Purple sea urchin *Strongylocentrotus purpuratus* and the fruit fly *Drosophila melanogaster*. Figure 9.2 shows the expected rooted four taxa tree derived from Fig. 9.1 and Fig. 9.3 shows the topologies of the three possible unrooted 4 taxa trees. By convention, Tree1 shows the generally accepted relationship among the four taxa while and Tree2 and 3 show the two alternatives. According to simple parsimony, the best tree is the tree that has the most phylogenetically informative characters (PIC) in its support. (This principle applies as well to weighted parsimony, maximum likelihoo, Bayesian, and protein distance methods though there are quantitative differences between these different approaches.) Let us assume that Tree1 represents the evolutionary history of the four taxa. Tree1 can then be supported by single changes that occur on the internal branch (refer to Fig. 9.3). Tree1 can also be supported by multiple (parallel or convergent) changes that occur on the distal branches that are homoplastic replacements. Thus, if Tree1 represents the actual history, then the number of phylogenetic informative characters (PIC) in its support (defined as N_1) will be those in which Sea urchin and human share one character and the fruit fly and jelly fish share another. N_1 will be determined by the sum of changes on the central branch and the homoplastic changes. There will also be PIC where the other two pairs of taxa share characters that can only arise by means of homoplastic changes on the distal branches. If the distal branches are relatively equal in length and the occurrence of homoplastic changes is randomly distributed, then we would expect to see the number of PIC due to homoplasy to be approximately equal, in which case $N_1 > N_2 = N_3$.

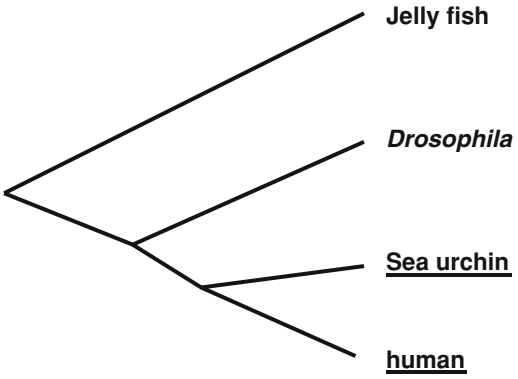


Fig. 9.2 Rooted 4 taxa tree. Shown are the four taxa that are used in the quartet partitioning. This is the generally accepted relationship among the four taxa showing the two deuterostomes—human and sea urchin to the exclusion of the protostome (drosophila) and jelly fish (cnidarian) outgroups

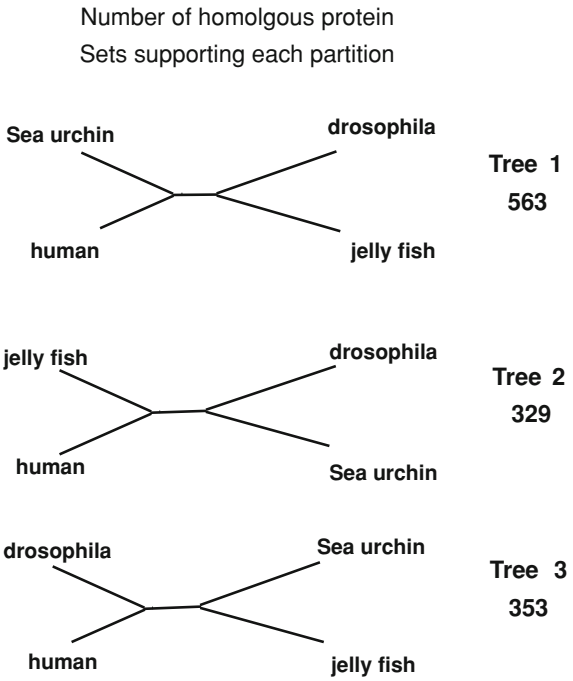


Fig. 9.3 Three unrooted four taxa topologies. Tree1 is the unrooted version of Fig. 9.2. Tree2 and Tree3 are the two remaining topologies. This defines the three topologies and the taxa used in quartet partitioning. The numbers correspond to the number of homologous protein sets that support each tree. The Tree3 seen here captures a relationship seen in Tree3 in our earlier paper (Syvanen and Ducore 2010)

Or more generally

$$N_i > N_j = N_k \tag{9.1}$$

where Tree i is the most parsimonious or the preferred tree.

A priori we can consider N_1 as putative support for Tree1, N_2 as support for Tree2 and N_3 as support for Tree3. By the principles of parsimony, the tree with the largest PIC in its support is the preferred one. Thus, the empirical finding of,

Table 9.1 Four taxa analysis of the protein sets in common between *H. sapiens*, *N. vectensis*, *S. purpuratus* and *D. melanogaster*

	N_1	N_2	N_3
(i) There 563 protein sets supporting Tree1			
PIC (total)	4,835	2,766	2,953
PIC (average)	8.5	4.9	5.2
Ratio	1.7	1.0	1.1
Chi sq (P) = 6 (0.1)			
(ii) There are 329 protein sets supporting Tree2			
PIC (total)	1,557	2,776	1,638
PIC (average)	4.7	8.4	4.9
Ratio	1.0	1.8	1.05
Chi sq (P) = 2 (0.5)			
(iii) There are 353 protein sets supporting Tree3			
Total	2,310	2,321	4,100
PIC (average)	6.5	6.6	11.6
Ratio	1.0	1.0	1.8
Chi sq (P) = 0.1 (0.95)			

The three sets were identified by the bootstrap partition. If more than 70 % of the bootstrap replicates supported Tree1, Tree2, or Tree3, then that protein set was assigned to that particular partition. Tree1 = (hu, su)(dr, cn), Tree2 = (hu, cn)(su, dr) and Tree3 = (hu, dr)(su, cn) as in Fig. 9.3. Chi square and (probability) give the results of the chi square that tests the distribution of N_1, N_2, and N_3 does not significantly deviate from the model $N_i > N_j = N_k$ (Eq. 9.1)

for example, $N_1 > N_2$ and N_3 is taken as evidence that Tree1 reflects the evolutionary history of the four taxa. In our approach, we applied the further restriction that Eq. 9.1 describes the PIC distribution and that deviation from this inequality raises questions about the consistency of the data. In four taxa analysis controls, we have shown that Eq. 9.1 holds reasonably well for those taxa that have undisputed relationships.

The current work with the four species shown in Fig. 9.2 begins by identifying a common set of proteins using Blast. Those proteins that are members of large gene families (i.e., copy numbers in excess of 10 in any of the four taxa) were excluded. This process identified about 2800, quartets that were aligned, submitted to parsimony analysis and the number of PICs in support of each tree determined as described in Methods. These sets will be referred to as "protein sets." The protein sets supporting alternative phylogenies are identified by determining the phylogeny for each of the approximately 2,800 protein sets and assessing bootstrap support for each tree. Each protein was submitted to a bootstrap analysis (200 replicates). Only those protein sets that had bootstrap support >70 % were included. There were only about 1,200 protein sets that significantly supported one of the three trees, the remaining 1,600 protein sets were excluded from further analysis. As summarized in Table 9.1, more protein sets support Tree1 (563

protein sets) than Tree2 (329 protein sets) or Tree3 (353 protein sets), nevertheless a large number of protein sets supported Tree2 and Tree3. The total number of PICs that support Tree1, Tree2, and Tree3, respectively, were combined and these combined sets support Eq. 9.1 (see Sect. 9.2.5 for further discussion of Table 9.1).

9.2.2 Nine Taxa Analysis

The four taxa partitions suggest potentially different phylogenies. However, there is not enough information to explain why. We therefore used the human sequence from each of the three partitions in separate Blast searches against a data base of nine taxa. This larger group of taxa includes two outgroups to the main metazoan cluster (the cnidarian, *N. vectensis* and the placozoan, *Trichoplax adherens*). The resulting trees can be rooted to these two outgroups so that the ancestral node for the Deuterostome/Protostome bifurcation can be identified. Increasing the number of taxa also allows one to assess whether or not "the taxon sampling artifact" is responsible for the incongruent trees revealed by the four taxa analysis (Lecointre et al. 1993; Matus et al. 2006; Dunn et al. 2008).

Let us first focus on the 9 taxa trees produced by the Tree1 and Tree3 supporting partitions. Figure 9.4 shows the topologies based on maximum likelihood analysis of the respective concatenated protein sets. There are a few salient points. Both trees preserve the 4 taxa topology that was found in Fig. 9.3. Hence, the phylogenetic information defining those two groups is not lost upon increasing the number of taxa. We can see that the two preselected outgroup taxa occupy an appropriate position in the tree, allowing us to infer a root. The Tree1 and Tree3 partitions produce clearly incongruent trees. Similar topologies were observed with parsimony and both Fitch and neighbor-joining analyses of protein distance matrices.

The two trees in Fig. 9.4 were submitted to a maximum likelihood analysis, and the log likelihood scores and standard deviations were recorded according to the Shimodaira and Hasegawa (1999) test to assess their differences. Table 9.2 shows the number of standard deviations separating the two datasets. The Tree1 partition significantly supports the Tree1 topology over Tree3 and vice versa.

The Tree2 partition gave conflicting results. This set seems to experience the taxon sampling artifact. Namely, the relative relationship of the four taxa defining the Tree2 partition changed as more taxa were added. When Tree2 partitions were submitted to the 9-taxa analysis, the four key taxa assumed topologies different from that which would be predicted by the 4 taxa analysis; some of these assumed a Tree1–like appearance and others a Tree3-like appearance (data not shown). This was not seen with the Tree1 and Tree3 partitions. These results indicate that there is considerable homoplasy in the character states for the Tree2 partition when additional taxa are added. No further effort to unravel this puzzle was made.

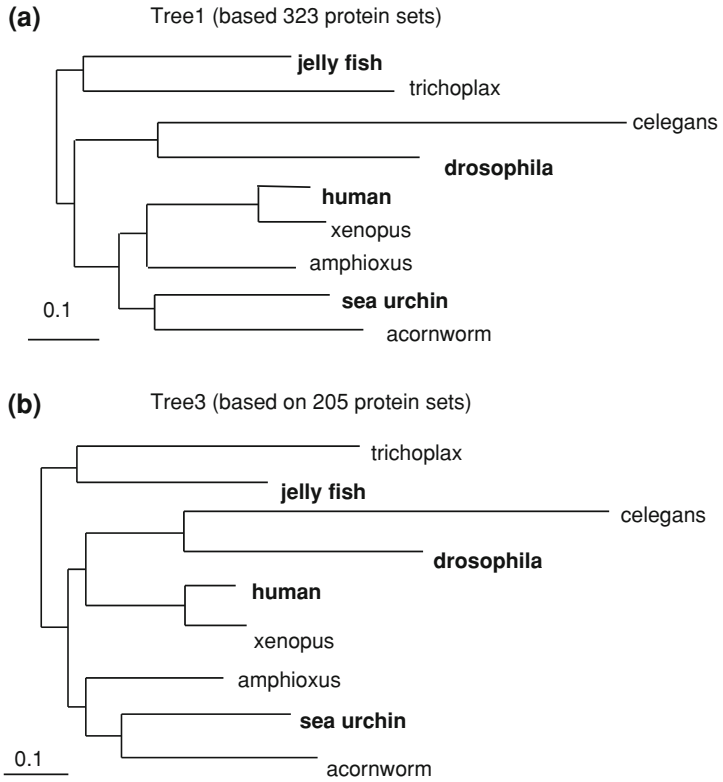


Fig. 9.4 **a** displays the tree using the Tree1 partition and **b** displays the tree using the Tree3 partition. All of the protein sets for each partition were concatenated into their respective file and the maximum likelihood trees were computed. The bold face taxa makes are those from Fig. 9.3

Table 9.2 Tree1 and Tree3 9-taxa topologies compared to the Tree1 and Tree3 character sets

Characters for:	Number std dev	
	From topology for:	
	Tree1	Tree3
Tree1	<0.1	15
Tree3	17	<0.1

Topology1 is from Fig. 9.4a and topology3 is from Fig. 9.4b and were submitted as user defined trees and analyzed by maximum likelihood. The number of standard deviations was determined using the Shimodaira and Hasegawa (1999) and Templeton test provided in Phylip

9.2.3 A Single Network Reconciling Tree1 and Tree3

Recently, phylogenetic networks have attracted attention as a useful analytical tool. If the underlying dataset contains conflicting signals that are due to reticulation (e.g., horizontal gene transfer or hybridization) a phylogenetic network may be more appropriate than a phylogenetic tree (reviewed in Huson and Scornavacca 2011).

To reconcile Tree1 and Tree3 from Fig 9.4 into a network, we used the algorithm HybridInterleave (Collin et al. 2013). This algorithm decomposes each of two input trees into a set of subtrees by deleting a minimum number of edges such that the resulting two sets are identical. In a subsequent step, the set of subtrees can then be used to reconstruct a network that explains all ancestral relationships given by the two input trees, and whose number of reticulation vertices (i.e., vertices with two incoming edges) is one less than the number of subtrees resulting from the decomposition step (Theorem 2 in Baroni et al. 2005).

Applying HybridInterleave to Tree1 and Tree3 from Fig. 9.4 results in the phylogenetic network depicted in Fig. 9.5. This network identifies the vertebrate as the hybrid clade while all other taxa have descended from the last common ancestor by vertical inheritance. In other words, if the vertebrate clade is removed from Tree1 and Tree3, the resulting 7-taxa trees are identical. Note that Fig. 9.5 depicts the unique phylogenetic network that results from applying HybridInterleave to Tree1 and Tree3. No other network with only one reticulation vertex can simultaneously explain Tree1 and Tree3. Based on this parsimony principle, the ancestor that gave rise to the vertebrate was a hybrid between an early protostome (or a sister group thereof) and a vertebrate ancestor that excludes the chordate amphioxus and the other two deuterostomes, the acorn worm and the sea urchin.

9.2.4 Temporal Patterns of Change

We made an effort to determine an age for the protostome deuterostome bifurcation shown in Fig. 9.4. These efforts were not fruitful because not only are the rates of change variable in the terminal branches of those trees, but variation in the rates of change in the internal branches seemed even larger. However, there is a feature seen in Fig. 9.4 that is noteworthy. We can see that the Human/xenopus clade is much closer to the root in Tree3 than it is in Tree1. Maximum likelihood distances on user defined trees can be in error if the trees are wrong. However, this truncated xenopus/human branch seen in the Tree3 protein set was also inferred using direct distances in pair-wise distance matrices, i.e., it is supported by the relative rate test. Vertebrate distances calculated from the Tree3 partition set are 18 % closer to the root of the tree than from the Tree1 partition set. This implies the presence of an ancestral lineage that is sister to the protostomes. Figure 9.6 shows the network that incorporates this inferred lineage.

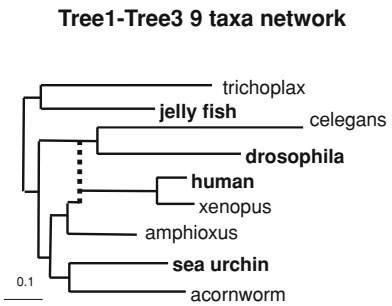


Fig. 9.5 The nine taxa network. The two trees in Fig. 9.2 were submitted to the HybridInterleave algorithm to find the most parsimonious network to resolve the conflicting trees

9.2.5 Alternative Explanations

Statistical Artifact

Let us consider the possibility that there is only a single class of protein sets supporting Tree1, but that the variance of N1, N2, and N3 is very high. According to this scenario, the partitioning method would have selected incorrect trees by chance. If there were a single distribution and it was based upon Tree1, we would expect to see three things not seen in the data.

First, if high variance were responsible for the Tree2 and Tree3 assignments, we would expect the number of PICs supporting the incorrect tree to be lower than those supporting Tree1 since small sample size is less reliable than large sample size. The data in Table 9.1 do not show this; in fact the number of PICs per protein set supporting Tree3 is greater than those supporting Tree1. Second protein sets from the Tree2 partition should show a distribution $N2 > N1 > N3$ and those from the Tree3 partition should show a distribution $N3 > N1 > N2$. This follows because in the first case we are selecting for $N2 > N1$ AND N3. While N1 may be a low outlier in some samples, N2 would be a high outlier in others. Since in either case N3 should remain normally distributed, we would expect $N1 > N3$. The same argument applies to the Tree3 protein set where we would expect $N1 > N2$. The data in Table 9.1 rules this out. Third, if a high variance were causing an incorrect assignment, we would expect that the Tree1 partition, as seen in the N1, N2, N3 distribution, would more robustly support its tree than would the Tree2 and Tree3 partitions. This is also not the case. As shown, the Ni:Nj:Nk distribution roughly equals 1.7:1:1 for each of the three partitions in Table 9.1.

Confusing Parology with Orthology

In our initial screen that identified protein sets, we selected for homology and did not distinguish orthology from parology. We tried to minimize this problem by eliminating large protein families in the original blast search. However, we should expect some paralogous families in the final dataset. These would arise if the last common ancestor of the four taxa contained multigene families but orthologues

were differentially lost in one of the taxa. In the original 4 taxa Blast searches for homologs, we used *H. sapiens* proteins as the query. The top score for each of the other three taxa were assembled as a protein set. We repeated the entire Blast search, once using the *N. vectensis* as the query and then again using *D. melanogaster* as the query. Basically, different Blast searches resulted in datasets (after the bootstrap selection for Tree1, Tree2, and Tree3 supporting protein sets) that yielded Tree1 and Tree3 9-taxa topologies similar to that seen in Fig. 9.4 (data not shown). To be sure, the number of Tree1 and Tree3 supporting protein sets resulting from these different searches are not the same. This is probably due to the fact that there are some sets composed of paralog/ortholog mixtures.

A second effort was made to minimize the parology/ortholog confusion. We edited the entire 1,200 protein set for duplicate unique identifying proteins and then removed those sets that had a protein found in more than one file. This truncated the number of protein sets but it did not change the relative support for Tree1, Tree2, and Tree3. Thus, there is no extreme bias toward selective loss in one taxa that could explain the tree incongruity in Fig. 9.4. The simplest explanation for the tree incongruity is a preexistence of two groups of protein sets with different evolutionary histories.

9.3 Discussion

The most straightforward explanation for the results presented here is that a primitive vertebrate ancestor, appearing after the split from the cephalochordates and tunicates, received an influx of genes from some unknown ancestor that is likely a sister group to modern protostomes. The size of the influx can only be approximated, but given the relative size of the partitions in Table 9.1, it appears that at least 20–30 % (if not more) of the modern vertebrate genes moved into the vertebrate lineage by this mechanism. If a single event is responsible, it is probably simplest to invoke a major hybridization between taxa that likely belonged to different phyla. Though entertaining such a big genetic upheaval may seem like a radical concept, there has been acceptance of the idea that a major genetic rearrangement occurred in an ancestor of the vertebrates that occurred after the cephalochordates and tunicates had diverged. This theory posits that there were two complete genome duplications during this period or that vertebrates evolved from a polyploid ancestor. Hughes and Friedman (2003) employed phylogenetic analysis of many duplicated genes to test this hypothesis. They found very little support for even a single duplication event, i.e., phylogenetic analysis revealed that duplicated regions of the chromosome diverged earlier than would be predicted by a simple genome duplication. If at least one of the major duplication events was the result of remote species hybridization, this pattern would be expected since that "duplication" event would be timed to the protostome-deuterostome speciation event, not to the hybridization event.

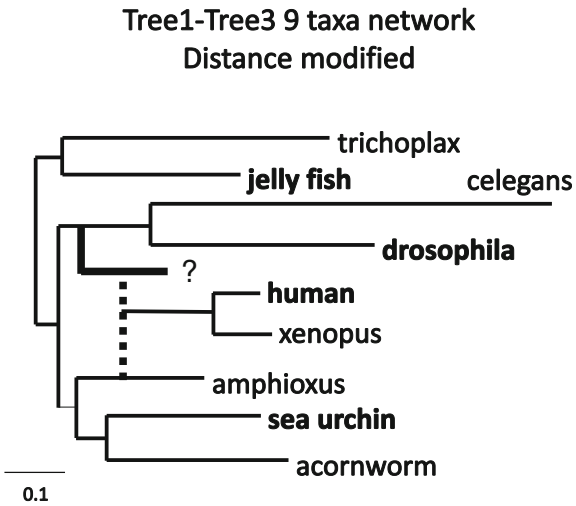


Fig. 9.6 Modified nine taxa network. A hypothesized internal branch was added to accommodate the molecular distance discrepancy seen in Fig. 9.4. The time to the last common vertebrate ancestor to in Tree1 (Fig. 9.4a) and Tree3 (Fig. 9.4b) must be the same. However, the molecular distance of the last common ancestor to is much shorter in Tree3 than in Tree1. This means the rate of evolution in an ancestral lineage for the Tree3 partition is much slower than is the rate for the Tree1 partition. Since the rate of evolution of the other protostome taxa seem to be even faster than the chordate lineages, a new unknown ancestral lineage is postulated

We have made some efforts to use molecular clock considerations to estimate the time of the hybridization event. This is not possible with any reasonable degree of precision, but we can see that there is considerable distance between the hypothesized influx of genes and the last common ancestor to the xenopus/human bifurcation; hence it appears quite possible that the event occurred early, probably before the Cambrian. We found that the 9-taxa Tree1 displays a much larger distance between the vertebrate LCA and the outgroup when compared to the LCA and outgroup distance in 9-taxa Tree3 (Fig. 9.4). This serendipitous result unexpectedly revealed properties of the donor ancestor. Namely, the donor ancestor experienced a relatively long period of evolution with an unusually slow molecular clock as compared to the extant taxa. This, we believe, reflects large differences in the rate of protein evolution among the lineages, including lineages in the internal branches. Further, these large differences in rates between ancestral lineages provide us with evidence that the evolutionary history of Tree3 partition proteins found in vertebrates followed a significantly different path than did the evolution of these proteins in the extant protostomes. Application of the HybridInterleave algorithm to the nine taxa Tree1 and Tree3 topologies identified the vertebrate as the hybrid clade. This unexpected pattern in rate also provides a second line of evidence supporting the hypothesis that the hybrid clade is the vertebrate.

The result shown in Fig. 9.5 is qualitatively similar to the result we published showing that the *C. intestinalis* evolved from a chordate-protostome hybrid

ancestor (Syvanen and Ducore 2010). Our earlier result cannot explain the current result. The four taxa partition sets that resulted from the *C. intestinalis*, vertebrate, *S. purpuratus, D. melanogaster* quartet behaved independently from the four taxa partitions produced in this study with the jelly fish, *H. sapiens*, *S. purpuratus, D. melanogaster* quartet. That is, the Tree1 partition from this study does not overlap with the Tree1 partition from the earlier study.

One can reasonably ask why such a major evolutionary event would have gone undetected until now. It should be stated that the differences between the protein sets found in partition 1 and partition 3 are not that large. For example, the average protein distances between these two partitions are very small compared to the variance between individual protein sets. Thus, for example, a distance distribution for the Tree1 and Tree3 partition sets, at first appearance, look the same. Additionally, there has not been much interest in looking for deep branch networks given that most work is devoted to finding a single tree (Fuchs et al. 2009; Delsuc 2009; Philip et al. 2005; Blair et al. 2002; DeSalle and Schierwater 2008) even when multiple trees are uncovered (Eitel et al. 2013; Nosenko et al. 2013).

A theory that posits major horizontal gene transfer early in metazoan history can explain two major observations: the taxon sampling paradox and leaf instability, which are two related phenomena that are a reflection of underlying homoplasy in the character data set. The classical "one true tree" theory deals with phenomena of this kind by assuming them to be unexplained noise. A theory that incorporates horizontal gene transfer can provide a mechanistic explanation. The results in this chapter also shed light on what has long been considered a paradox. The fossil record supports the notion that the modern metazoan phyla radiated from a single point in time. However, modern genomics has established that multiple and varied ancestral animals preceded and contributed to the post Cambrian explosion, and considerable parallelism in morphological evolution is evident. A theory of evolution incorporating horizontal gene transfer can also easily explain that apparent paradox.

9.4 Materials and Methods

A group of 3,800 protein sequences from the Human genome sequence were used as query sequences in Blast searches. These sequences were selected from ca 25,000 human proteins on the basis of having homologs in a variety of other metazoa and also belonging to gene families with a copy number less than 10. Searches were made against a database consisting of the protein sequences obtained from the genome projects for the following metazoans: *H. sapiens* (Human Genome Resources 2010), *Xenopus laevis* (JGI 2009) the sea urchin *Strongylocentrotus purpuratus* (Sea Urchin et al. 2006) the fruit fly *D. melanogaster* (Celniker et al. 2002) and the round worm *Caenorhabditis elegans* (*C. elegans* 1998) the amphioxus *Branchiostoma floridae* (Nicholas 2008) the cnidarian *Nematostella vectensis* (Sullivan et al. 2006), and the placozoan

Trichoplax adherens (Srivastava et al. 2008) and the acorn worm *Saccoglossus kowalevskii* (acorn worm). An expectation score of less than 10^{-13} was used in all cases. These Blast results were screened such that each contained at least one homolog for the particular 4 taxa or 9 taxa analysis. For those proteins that had multiple listings for the same taxa, the protein with the smallest expectation value was used.

Sequences from each output file were recovered, and multisequence alignments were performed using Clustal (Thompson et al. 1994) and then gaps were deleted with the sequence editor Gblocks (Castresana 2000). Phylogenetic analysis was performed using the Phylip suite of programs (Felsenstein 2005). Four different types of trees were determined. Simple parsimony, maximum likelihood, nearest neighbor, and Fitch distance trees were determined as noted. For tree and molecular clock estimations protein distances were calculated after concatenating the protein sets for each partition. The Jones, Thornton, Taylor distance matrix (Jones et al. 1992) was used in the distance and maximum likelihood methods. In preliminary screens of the protein sets it was shown that distances up to 2.5 changes per residue were linear with time of divergence (data not shown), and those protein sets containing distances in excess of 2.5 were removed from further consideration. The phylogenetic maximum likelihood program proml was used to calculate log likelihood scores that uses the Shimodaira-Hasegawa test (1999). Programs within Phylip were also used to perform the bootstrap procedure. Treeview (Page 1996) was used for tree visualization. All computations were performed on a standard pc with a Linux OS and data was processed using shell script files, Perl scripts and standard spread sheets. The HybridInterleave algorithm was used transform two incongruent phylogenetic trees into a single phylogenetic network (Collins et al. 2011).

The number of phylogenetic informative characters (N) that supports tree i is $N_i = (\text{pic} - 2T_i + T_j + T_k)/3$ where PIC is the total of number of PICs and T is the total length of the parsimony tree in units of unweighted amino acid differences. In a four taxa tree the only PIC are those in which two taxa share one amino acid and the other two share another.

References

Acorn worm site. ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Skowalevskii/fasta

Baroni M, Grünewald S, Moulton V, Semple C (2005) Bounding the number of hybridisation events for a consistent evolutionary history. J Math Biol 51:171–182

Blair JE, Blair Hedges S (2005) Molecular phylogeny and divergence times of deuterostome animals. Mol Biol Evol 22:2275–2284

Blair JE, Ikeo K, Gojobori T, Hedges SB (2002) The evolutionary position of nematodes. BMC Evol Biol 8(2):7

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 11:2012–2018

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552
Celniker et al (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol 3:1–0079
Collins J, Linz S, Semple C (2011) Quantifying hybridization in realistic time. J Comput Biol 18:1305–1318
Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968
DeSalle R, Schierwater B (2008) An even newer animal phylogeny. BioEssays 30:1043–1047
Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H (2004) The timing of eukaryotic evolution. Proc Natl Acad Sci USA 101:15386–15391
Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sørensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749
Eitel M, Osigus HJ, DeSalle R, Schierwater B (2013) Global diversity of the Placozoa. PLoS One 8(4):e57131
Felsenstein J (2005) PHYLIP (Phylogeny inference package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle
Fuchs J, Obst M, Sundberg P (2009) The first comprehensive molecular phylogeny of Bryozoa (Ectoprocta) based on combined analyses of nuclear and mitochondrial genes. Mol Phylogenet Evol 52:225–233
Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162
Gauthier O, Lapointe F (2007) Hybrids and phylogenetics revisited: a statistical test of hybridization using quartets. Syst Bot 32:8–15
Hillis DM, Huelsenbeck JP (1992) Signal, noise, and reliability in molecular phylogenetic analyses. J Hered 83:189–195
Hughes AL, Friedman R (2003) 2R or not 2R: testing hypotheses of genome duplication in early vertebrates. J Struct Funct Genomics 3:85–93
Human Genome Resources 2010. http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/
Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267
Huson DH, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. Genome Biol Evol 3:23–35
JGI *X. tropicalis* genome assembly (2009). http://genome.jgi-psf.org/Xentr4/Xentr4.home.html
Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282
Lecointre G, Philippe H, Vân Lê HL, Le Guyader H (1993) Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol 2:205–224
Matus DQ, Copley RR, Dunn CW, Hejnol A, Eccleston H, Halanych KM, Martindale MQ, Telford MJ (2006) Broad taxon and gene sampling indicate that chaetognaths are protostomes. Curr Biol 8:R575–R576
Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WE, Nickel M, Schierwater B, Vacelet J, Wiens M, Wörheide G (2013) Deep metazoan phylogeny: when different genes tell different stories. Phylogenet Evol 67:223–233
Osigus HJ, Eitel M, Bernt M, Donath A, Schierwater B (2013) Mitogenomics at the base of Metazoa. Mol Phylogenet Evol 69:339–351
Page RDM (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357–358
Peterson KJ, Cotton JA, Gehling JG, Pisani D (2008) The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos Trans R Soc Lond B Biol Sci 363:1435–1443

Philip GK, Creevey CJ, McInerney JO (2005) The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 22:1175–1184

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19:706–712

Putnam NH, T Butts, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguch E, Terry A et al (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071

Sea Urchin Genome Sequencing Consortium, Sodergren E et al (2006) The Genome of the sea urchin Strongylocentrotus purpuratus. Science 314:941–952

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

Srivastava, M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J (2008) The Trichoplax genome and the nature of placozoans. Nature 454:955–960

Sullivan JC, Ryan JF, Watson JA, Webb J, Mullikin JC, Rokhsar D, Finnerty JR (2006) StellaBase: the Nematostella vectensis genomics database. Nucleic Acids Res 1:34

Syvanen M (1985) Cross-species gene transfer: implications for a new theory of evolution. J Theor Biol 112:333–343

Syvanen M (2002) On the occurence of horizontal gene transfer among an arbitrarily chosen group of 26 Genes. J Mol Evol 54:258–266

Syvanen M (2012) Evolutionary implications of horizontal gene transfer. Ann Rev Genet 46:341–358

Syvanen M, Ducore J (2010) Whole genome comparisons reveals a possible chimeric origin for a major metazoan assemblage source. J Biol Syst 18:261–275

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Wray GA, Levinton JS, Shapiro LH (1996) Molecular evidence for deep Precambrian divergences among metazoan phyla. Science 274:568–573

Chapter 10
Evidence for Ancient Horizontal Gene Acquisitions in Bdelloid Rotifers of the Genus *Adineta*

Boris Hespeels, Jean-François Flot, Alessandro Derzelle and Karine Van Doninck

Abstract Until recently, obligate asexuality was often considered an evolutionary dead end. However, recent advances suggest that conventional sexual reproduction, defined as the alternation of meiosis and fertilization, is not the only sustainable eukaryotic lifestyle. Moreover, different modes of asexual reproduction are observed in nature, raising the question of the diverse mechanisms responsible for the long-term survival and adaptation of strict asexuals. One possible way to study the molecular-genetic consequences of the loss of meiotic recombination is to scrutinize the genomes of asexuals of ancient and more recent origins. The first genome draft of an ancient asexual species, the bdelloid rotifer *Adineta vaga,* was recently made available, revealing a peculiar genomic structure in which allelic regions were massively rearranged and sometimes found on the same chromosome. Such genome organization devoid of homologous chromosomes appears incompatible with meiotic pairing and segregation, and represents therefore a compelling genomic signature of asexuality. Besides, the genome of *A. vaga* contains around 8 % of genes of apparent nonmetazoan origin, a percentage much higher than observed in most eukaryotes. Interestingly, a similar percentage of genes of nonmetazoan origin was independently inferred from a large-scale transcriptome analysis of the bdelloid rotifer *Adineta ricciae*. In this chapter, we conducted a comparative study between these two closely related species using reciprocal best blast hits, followed by functional annotation using the GOANNA pipeline. Around 10 % of all the orthologs identified between the two species were putatively acquired by horizontal gene transfer and lots of them were associated to hydrolases (18 %) and oxidoreductases (16 %) functions. We hypothesize that these acquisitions may have helped bdelloids to adapt to multiple food sources and

B. Hespeels · A. Derzelle · K. Van Doninck (✉)
Laboratory of Evolutionary Genetics and Ecology, URBE, Department of Biology, University of Namur, 5000 Namur, Belgium
e-mail: karine.vandoninck@unamur.be

J.-F. Flot
Max Planck Institute for Dynamics and Self-Organization, Bunsenstraße 10, 37073 Göttingen, Germany

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_10,

to develop enhanced resistance to desiccation. Furthermore, comparisons with sequences available for the monogonont rotifer *Brachionus plicatilis* suggest that some nonmetazoan genes were acquired by rotifers before the separation of bdelloids and monogononts.

10.1 Introduction

Sexuality is considered the dominant mode of reproduction in the animal kingdom. The supremacy of sex among metazoans is usually explained by invoking its supposed long-term evolutionary advantages. Sex brings together genetic material from different individuals during fertilization, and shuffles it through genetic recombination during meiosis. This mixing provides novel allelic combinations that subsequently undergo selection, enhancing adaptation to the environment (Maynard Smith 1978; Lesbarrères 2011; Lodé 2011). Asexuals, on the contrary, lack meiosis and are therefore expected to exhibit less genetic diversity (hence lower adaptation rates). Moreover, genomes that do not experience sexual recombination have been described to fall prey to an irreversible, stochastic accumulation of deleterious mutations known as Muller's ratchet (Muller 1932; Felsenstein 1974). It is mainly because of these two reasons (low genetic diversity and accumulation of deleterious mutations) that obligate asexuality is often considered an "evolutionary dead end". Asexual populations of the model crustacean *Daphnia pulex* illustrate well the rapid evolutionary turnover of asexual animals. In *D. pulex*, obligate asexuality arises when a male transmits a dominant meiosis-suppressing allele to its offspring (Lynch et al. 2008). Interestingly, all current asexual populations of *D. pulex* bear "a meiosis-suppressing element" that originated only 1250 years ago, and the populations themselves appear even younger (Tucker et al. 2013). This suggests that asexual *Daphnia* populations do not survive in the long-term and appear doomed to extinction.

Nevertheless, some asexual metazoans have apparently been highly successful for millions of years: the so-called 'ancient asexuals'. The most notorious ones are darwinulid ostracods, oribatid mites, *Meloidogyne* root-knot nematodes (RKNs) and bdelloid rotifers (Danchin et al. 2011). The latter group, comprising more than 400 morphospecies, has apparently been able to diversify without sexual reproduction and was labeled for this reason by Maynard Smith as "something of an evolutionary scandal"(Maynard Smith 1986). Beside these asexual metazoans, many asexuals have also been inventoried in other eukaryotic groups, adding up to 20 % of all fungi for instance (Judson and Normark 1996; Seidl and Thomma 2014). These examples suggest that the alternation of meiosis and fertilization events is not the only sustainable eukaryotic life cycle. This implies the existence of alternative mechanisms that may prevent mutation accumulation and generate genetic diversity on the long term, thus making up for the absence of sex.

10.2 Genomic Insights into Ameiotic Evolution

One possible way to study the molecular-genetic consequences of the loss of meiotic recombination is to tap the genomes of asexuals, both of ancient and recent origins.

A comparative whole-genome analysis of sexual and asexual lineages of *D. pulex* was recently carried out (Tucker et al. 2013). This study, as mentioned in the introduction, indicates that asexual *D. pulex* lineages are much younger than sexual ones. Interestingly, heterozygosity does not seem to increase in asexual *Daphnia* as expected under the so-called Meselson effect, i.e., the accumulation of differences between alleles that do not experience recombination (Birky 1996). On the contrary, gene conversion and gene deletion events appear to be more common than point mutations. *Timema*, another interesting example, is a genus of plant-feeding insects among which at least five independently derived asexual lineages have been described (Schwander et al. 2011; Henry et al. 2012). The estimated ages of these asexual lineages range from 100,000 to over 1 million years. The occurrence of both sexual and asexual species in *Timema* makes it an ideal model to study the origin and evolution of asexuality over time. Comparing allelic variation between two *Timema* sister species revealed a higher level of nuclear allelic divergence in the asexual line (Schwander et al. 2011) consistent with the Meselson effect. Interestingly, in a few asexual *Timema* groups, only one allelic copy of each gene was found to accumulate deleterious mutations. Therefore, these asexuals appear to have the capability to make up for the functional loss of one copy from each of their genes (Schwander et al. 2011), thereby evolving toward haploidy. This trend is also visible at chromosome scale in one species (*Timema genevievae*) whose karyotype is devoid of pairs of homologous chromosomes (Schwander and Crespi 2009). The evolution towards haploidy observed in asexual *Timema* species contrasts with the hypothesis that polyploidy may be favored as it slows down Muller's ratchet: hence, studying several independent examples of asexuals appears necessary to understand all the possible genomic consequences of a switch from sexual towards asexual reproduction.

The genome sequences of sexual (*Meloidogyne hapla*) and asexual (*Meloidogyne incognita*) species of RKNs have provided a powerful system to improve our understanding of these consequences. A striking difference is that the size of the genomic assembly of *M. incognita* is 86 Mb, almost twice the haploid size suggested by flow cytometry (50 Mb), whereas for its sexual relative *M. hapla* the assembly size (54 Mb) agree with the haploid size estimate (Abad et al. 2008). This is because 64 % of the genome assembly of *M. incognita* appears to be made up of highly divergent allelic regions (with an average nucleotide divergence between two copies of 7 %) that were assembled separately (Castagnone-Sereno and Danchin 2014). Despite this separate assembly of allelic regions, there were no pairs of large scaffolds completely colinear to each other (Castagnone-Sereno and Danchin 2014). Such absence, first described in the bdelloid rotifer *A. vaga* (Flot et al. 2013; see below), is considered a signature of long-term ameiotic evolution,

and is consistent with previous reports of a diversity of chromosomes numbers in *M. incognita* (Triantaphyllou 1981). Initially, the high level of nucleotide divergence observed in *M. incognita* was considered an instance of the Meselson effet (Castagnone-Sereno and Danchin 2014), which may result in alleles becoming functionally divergent (as paralogs do). However, subsequent phylogenetic analyses revealed that alleles did not cluster according to morphological species, suggesting that asexual RKN lineages are actually allopolyploids (Lunt 2008; Castagnone-Sereno and Danchin 2014).

Although the first bdelloid genome became available 5 years after *M. incognita*, it was this genome that provided the most thorough breakthrough in our understanding of the genomic peculiarities of asexual animals. Bdelloid rotifers comprise more than 460 morphospecies, among which no sign of sexuality has ever been observed: there are no male organs, and cytological studies by Hsu (1956a, b) on two distinct bdelloid species reported that oogenesis proceeds in a mitotic way, without meiotic pairing nor reduction in chromosome numbers. However, these observations were not sufficient to exclude that bdelloid rotifers engage in some rare, cryptic mode of sexual reproduction. Indeed, their sister group the monogononts are cyclical parthenogens that only produce males when the environmental conditions start to deteriorate. The recent publication of the genome draft of a bdelloid rotifer species, *A. vaga,* has brought strong new support to the hypothesis of their asexuality by revealing a peculiar genomic structure in which allelic regions are massively rearranged and sometimes found on the same chromosome, in most cases in a palindromic fashion (Flot et al. 2013). The absence of homologous chromosomes is incompatible with meiotic pairing (as mentioned above in the case of *M. incognita*), whereas the physical link between allelic regions co-occurring on the same chromosome precludes meiotic segregation: hence, it appears very unlikely that this bdelloid lineage is able to perform meiosis. The average divergence between allelic regions in *A. vaga* is 4.4 %, less then in *M. incognita*. Although relatively high among metazoans, this average heterozygosity fall within the upper range observed for sexually reproducing species (Leffler et al. 2012). About one fifth of the genes of *A. vaga* are present in four copies, indicating that it is an ancient tetraploid that has lost already many of its duplicated genes.

In addition to shedding light on the peculiar genome structure of *A. vaga*, the sequencing of the genome of this species also revealed genetic signatures of frequent gene conversion between homologous regions (Flot et al. 2013). The inferred lengths of the conversion tracks ranged from tens of bases up to hundred thousand bases (in which case the near-identical copies collapsed during genome assembly, resulting in regions of double coverage depth). Gene conversion was hypothesized to play an important role in the long-term survival of bdelloids in the absence of sexual recombination: first, because gene conversion slows down Muller's ratchet (Connallon and Clark 2010; Flot et al. 2013); and second, because gene conversion either exposes recessive deleterious mutations to selection by turning them into homozygous state or removes them by overwriting them with the other allele, in both cases decreasing the mutational load of the genome (Khakhlova and Bock 2006; Flot et al. 2013).

10.3 Mechanisms Enhancing Genomic Plasticity

The genomic data available on asexual animal lineages suggest several mechanisms that can enhance genomic plasticity in the absence of meiotic recombination. Both the genomes of *A. vaga* and *M. incognita* contain numerous synteny breakpoints (Flot et al. 2013; Castagnone-Sereno and Danchin 2014), suggesting a dynamic genome structure: such genomic plasticity may be one of the keys to understand how asexuals adapt to their environment. In the asexual phytopathogenic fungus *Verticillium dahliae* a variety of large genomic rearrangements have also been observed, in contrast with the low degree of genome-wide nucleotide diversity of this species (de Jonge et al. 2013). These rearrangements lead to the individualization of lineage-specific genomic regions enriched in genes involved in adaptation to the host plant, thereby increasing the virulence of this asexual fungus. These results are the first compelling evidence that genome rearrangements facilitate adaptation in asexual organisms, as was also hypothesized to be the case for some cancerous lineages (chromothripsis; cf. Stephens et al. 2011).

Transposable elements (TEs) have been proposed as important mediators of genomic plasticity (Seidl and Thomma 2014). TEs are genomic DNA segments that are able to move around a genome in a copy/paste or cut-and-paste fashion, inducing mutations and chromosomal rearrangements (Gladyshev and Arkhipova 2010). Multiple major chromosomal rearrangements induced by TEs have been described in plants (Zhang et al. 2011). In *Saccharomyces cerevisiae*, TEs have been described to induce rearrangements allowing fast adaptation to the environment (Crombach and Hogeweg 2007). Interestingly, variation in genomic structure through chromosomal rearrangements was associated to increased fitness during asexual growth of *Schizosaccharomyces pombe* (Avelar et al. 2012). In the asexual fungus *V. dahliae*, rearranged portions of chromosomes are also mostly flanked by retrotransposons and repetitive sequences (de Jonge et al. 2013). Comparative genomic analyses between sexual and asexual nematodes detected a higher proportion of TEs and repetitive sequences in *M. incognita* (Abad et al. 2008) than in its closest sexual relative *M. hapla* (36 vs. 12 %, respectively). Interestingly, a putative full-length Tm1 transposase has been detected in the genomes of the mitotic parthenogenetic species *M. javanica* and *M. incognita* (Gross and Williamson 2011). By contrast, no functional Tm1 transposase sequence was detected in *M. hapla*. These observations suggest that functional TEs are present in the genomes of asexual nematodes and may play a role in their genomic plasticity. However, the maintenance of a high fraction of TEs in an asexual genome also carries a risk of unrestrained genomic expansion that could cause the extinction of the lineage (Arkhipova and Meselson 2005). Indeed Arkhipova and Meselson (2005) hypothesized that sexual reproduction can limit the expansion of TEs through ectopic crossing-over and homologous recombination. In contrast, in the absence of sex, TEs are hypothesized to multiply indefinitely, leading to population extinction (Arkhipova and Meselson 2005). Ancient asexuals are therefore likely to possess genome defense mechanisms that prevent TE

expansion. A recent comparison of TE content in sexual and asexual wasps suggests that reality is even more complex than the theoretical considerations above (Kraaijeveld et al. 2012). In this study, TE content was compared in a sexual lineage of the parasitoid wasp *Leptopilina clavipes* and in another lineage in which *Wolbachia* bacteria had induced parthenogenesis. Despite the presence of TEs in both sexual and asexual wasp lineages, there was no evidence of an overall increase in copy number for all TEs in the asexual populations. However, one group of TEs (the *gypsy*-like LTR elements) was overrepresented in the asexual lineage. This could be caused by the manipulation of the host genome by *Wolbachia* rather than by a direct impact of asexuality (Kraaijeveld et al. 2012). Only 3 % of the genome of the asexual bdelloid species *A. vaga* is made up of TEs, despite the high diversity of bdelloid TE families (255 in total), each of which is present in very low copy numbers (Flot et al. 2013). Moreover, most detected TEs appear to be recent arrivals and the protein families involved in the epigenetic silencing of TEs are substantially expanded in the genome of *A. vaga*, suggesting that incoming TEs are quickly reduced to silence (Flot et al. 2013; Arkhipova et al. 2013). Despite their low abundance, a role of TEs in promoting copy number variation in *A. vaga* has been suggested based on the observation of expanded gene families surrounded by TE footprints (Arkhipova et al. 2013).

In bdelloid rotifers, exposure to frequent desiccation events in their temporary habitats has been hypothesized to play a role in promoting structural genomic rearrangements and gene conversion. Bdelloids are highly adapted to semi-terrestrial environments: at any stage of their life cycle, when their habitat dries out, they can enter a metabolically quiescent state of desiccation for a prolonged period of time. Recently, it was demonstrated that desiccated *A. vaga* individuals accumulate DNA double-strand breaks (DSBs) through time and start repairing them once they become rehydrated (Hespeels et al. 2014). Interestingly, desiccation-induced DNA DSBs may promote gene conversion through mitotic recombination during DSB repair, which could in turn to prevent the accumulation of deleterious mutations in this asexual lineage (Flot et al. 2013). Moreover, the repair of DNA DSBs induced by desiccation may promote genome rearrangements and copy number variation. Copy number variation could be more frequent than point mutation, as in *C. elegans* where it is two orders of magnitude higher (Lipinski et al. 2014), and may therefore play an important role in generating genome variation. As a result of such duplications, some facultative asexual lineages of aphids and *Daphnia* present numerous duplications or even expansion of gene families, resulting in a number of gene more than twice the arthropod average (International Aphid Genomics Consortium 2010; Colbourne et al. 2011). Genes present in multiple copies can diverge and eventually acquire novel functions, enhancing in turn genomic plasticity (Castagnone-Sereno and Danchin 2014). Further studies on the evolutionary dynamics of gene conversion, gene duplications, and functional divergence in asexual lineages will be required to find out whether these mechanisms enhancing genomic plasticity may be sufficient to allow the adaptation and long-term persistence of asexual animals.

10.4 Horizontal Gene Transfers in Asexual Metazoans

Horizontal gene transfer (HGT) used to be considered relevant to prokaryotes only. Indeed, there seems to be some important barriers to HGT in eukaryotes. First, foreign DNA needs to pass both the eukaryotic cell membrane and the nuclear envelope. Second, metazoan germline cells are usually segregated from the rest of the body and not in contact with the outside world, reducing the likelihood that genes acquired from the environment are passed down to the next generation. Third, in contrast to prokaryotes where three distinct pathways of HGT have been well-characterized (conjugation, transformation, and transduction; Ochman et al. 2000), no general mechanisms allowing HGT has been described in eukaryotes.

The recent increases in analytical power and in the number of sequenced eukaryotic genomes have dramatically changed this view. Although only a few metazoans have been screened for HGTs, the percentage of genes of suspected HGT origin in their genome was consistently around 0.5–1 %, and much higher in a few species (Table 10.1). Most of these genes seem to be functional and encode proteins involved in metabolic processes. One of the sources of HGTs to eukaryotic organisms are their symbionts. For instance, the fruit fly *Drosophila ananassae* has integrated nearly the whole genome of its bacterial endosymbiont *Wolbachia* (living in close association with the host gonadic tissues): in total, 8 % of the genes of *D. ananassae* originated from *Wolbachia* (Hotopp et al. 2007). However, the functional significance of these integrated *Wolbachia* genes is not known. HGTs from *Wolbachia* genes into the genome of its host were also observed in other insects and nematodes known to be infected by this bacterial genus. Despite these examples, integration of symbiont DNA is not the only source of HGT: for instance, the aphid *Acyrthosiphon pisum* and the cnidarian *Hydra magnipapillata* harbor nonmetazoan genes that did not originate from their identified symbionts (Nikoh et al. 2010; Chapman et al. 2010).

Interestingly, numerous HGT events have been reported in nematodes. These foreign transfers have apparently played a key role in the acquisition of parasitic capacities by some nematode species. In *Bursaphelenchus xylophilus*, for instance, the transferred genes seem to be involved in adaptation to pine-tree parasitism (Kikuchi et al. 2011). *Pristionchus pacificus*, a self-fertilizing nematode, has a complex life cycle inside a beetle host: the first part of the larval development occurs when the host is alive, then development stops and resumes once the host is dead. The decaying body of the beetle becomes a complex ecosystem containing bacteria, fungi, and unicellular eukaryotes, providing ample sources for HGTs. Indeed, 2.1 % of the genes of *P. pacificus* are putative HGTs from diverse phyla. Interestingly, a strong enrichment in genes of apparent insect origin was reported, suggesting that the intimate physical contact between the nematode and its host promotes HGT. Like *A. vaga*, *P. pacificus* presents an enrichment in TEs in the genomic neighborhood of horizontally transferred genes (Rödelsperger and Sommer 2011). Multiple HGTs have also been reported in RKNs of the genus *Meloidogyne* and are suspected to have favored the emergence of plant parasitism

Table 10.1 List of metazoan species screened for horizontally transferred genes

Species name	%HGT	%TE	Discovery method	Reproductive mode	Reference
Acyrthosiphon pisum (pea aphid)	0.06	40.2	Phylogeny	Sexual/asexual	Nikoh et al. (2010); Moran and Jarvik (2010)
Bursaphelenchus xylophilus (nematode)	0.13	/	Phylogeny	Sexual	Kikuchi et al. (2011)
Bombyx mori (silkmoth)	0.29	45	Phylogeny	Sexual	Zhu et al. (2011)
Hydra magnipapillata (hydrozoan)	0.36	57	Phylogeny	Sexual/asexual	Chapman et al. (2010)
Drosophila melanogaster (fruit fly)	0.6	15	h_U	Sexual	Boschetti et al. (2012)
Ciona intestinalis (tunicate)	0.66	/	Phylogeny	Sexual	Ni et al. (2012)
Brachionus plicatilis (monogonont rotifer)	1.8	/	h_U	Sexual/asexual	Boschetti et al. (2012)
Caenorhabditis elegans (nematode)	1.8	12	h_U	Sexual/asexual	Boschetti et al. (2012)
Pristionchus pacificus (nematode)	2.10	<15 %	Codon usage, phylogeny	Self-fertilization	Rödelsperger and Sommer (2011)
Meloidogyne incognita (nematode)	3.19	6.3	Phylogeny	Asexual	Paganini et al. (2012)
Adineta vaga (bdelloid rotifer)	8	3	AI	Ancient asexual	Flot et al. (2013)
Adineta ricciae (bdelloid rotifer)	9.6	/	h_U	Ancient asexual	Boschetti et al. (2012)
Drosophila ananassae (fruit fly)	8	/	PCR	Sexual	Hotopp et al. (2007)

Most studies used phylogenies to uncover HGT: in this approach, horizontally acquired genes are detected because they show up in a different, unexpected part of the phylogeny. Other commonly used methods include the HGT index h_U, equal to the difference between the highest nonmetazoan and the highest metazoan bitscores (Boschetti et al. 2012), and the "alien index" (AI), computed as the log-ratio of the BLAST E-values for the best metazoan hit and the best nonmetazoan hit for a given gene (Gladyshev et al. 2008). In the fruit fly *D. ananassae*, *Wolbachia* sequences were detected by direct PCR amplification (Hotopp et al. 2007)

within this genus (Danchin et al. 2010; Haegeman et al. 2011). HGTs are found in both sexual and asexual *Meloidogyne* species but the most virulent RKNs are the asexual ones (*M. incognita*, *M. javanica,* and *M. arenaria*, together responsible for 90 % of worldwide agricultural plant damage). A majority of the HGTs in *M. incognita* are enzymes involved in the degradation or modification of plant cell walls and were acquired from bacteria sharing the same ecological niche as RKNs. Compared with other genes, HGTs in *M. incognita* are characterized by their higher degree of autonomy, i.e., they do not require synergy with other genes to yield a function and are therefore more easily transferable horizontally than genes involved for instance in the regulation of metabolic processes. In addition, many of the genes acquired horizontally apparently experienced duplication after their acquisition. This process of gene duplications facilitates the emergence of multigene families where each copy can yield a better or divergent function through neo-functionalization (Paganini et al. 2012).

In bdelloid rotifers, a first transcriptomic study of the species *A. ricciae* submitted to desiccation revealed that around 10 % of the genes expressed under hydrated or desiccated conditions were potentially acquired by HGT; by contrast, in monogonont rotifers, the percentage of HGT is only 1.8 % (Table 10.1) (Boschetti et al. 2012). Most of these acquired sequences (80 %) appear to be involved in enzymatic reactions such as toxin degradation and antioxidant production. Similarly to *M. incognita*, these data are consistent with the hypothesis that genes that can act independently are more easily transferred and/or retained in the recipient genome than genes integrated in a pathway and requiring other proteins to function. Multiple independent acquisitions from up to 533 distinct source organisms (bacteria, fungi, plants,…) are suspected to have happened in *A. ricciae* (Boschetti et al. 2012). A similar level of abundance of HGTs was observed in the genome of *A. vaga*, where an estimated 8 % of the geneset present strong signatures of nonmetazoan origin. As 20 % of these foreign genes are present in quartets, it was hypothesized that these transfers occurred before the establishment of tetraploidy and therefore before the separation of the bdelloid families (Hur et al. 2009; Flot et al. 2013). However, 60 hypothetical genes acquired horizontally in *A. vaga* had no intron, were present in one copy, and presented a GC content markedly different from the genome average, suggesting that they were acquired quite recently (Flot et al. 2013). Hence it appears that horizontal gene acquisition in bdelloid rotifers is still ongoing.

The availability of both a transcriptomic dataset of *A. ricciae* and a genome draft of *A. vaga* allowed us to use for the first time a comparative analysis to answer some key questions related to HGT in bdelloids. How many genes are shared by both species? What are their functions? How many orthologs shared between the two species were potentially acquired by HGT prior to their divergence? Can they inform us about past events in the evolutionary history of the genus *Adineta* and of rotifers in general?

10.5 Looking for Orthologous Genes Between *A. vaga* and *A. ricciae*: Genomic Evidence for Ancient Horizontal Gene Transfers in Bdelloids

Orthologs, a special class of homologs, are defined as genes originating in a common ancestor species and having evolved independently in different species following speciation. Orthologs usually share similar functions, in contrast to paralogs (homologs originating from a gene duplication event) that often diverge in function following duplication (Fitch 1970; see Koonin 2005 for review). Here, we focus our comparative analysis on orthologs of *A. vaca* and *A. ricciae* of putative nonmetazoan origin, in order to determine whether HGT acquisition occurred before or after the separation of these two lineages and how these genes became apparently "domesticated" in bdelloids.

The bdelloid genus *Adineta* is composed of 14 morphological species (Segers et al. 2007). However, molecular data suggest that cryptic species are frequent within this genus (Fontaneto et al. 2009, 2010). Here, we compared the draft genome of one lineage of the morphospecies *A. vaga* originating from Matthew Meselson's laboratory at Harvard University (Flot et al. 2013) with cDNA libraries from *A. ricciae* (Boschetti et al. 2012). The latter species, discovered in a dry pond in Australia, presents a morphology close to *A. vaga* with a few differences: *A. ricciae* has eyes in frontal position that are absent in *A. vaga* and the two species have different numbers of mastax teeth (Segers and Shiel 2005). Both species were selected for high-throughput sequencing because they are easy to culture in vitro (in contrast to most bdelloids collected from the wild) and are therefore amenable to molecular biology analyses requiring thousands of individuals.

Reciprocal best blast hits (RBBHs), also known as best bidirectional hits (Hulsen et al. 2006), provide a simple and reliable method to detect orthologs between two species (Altenhoff and Dessimoz 2009). This approach considers that two genes are orthologous if they are each other's best hit when aligning the geneset of one species against the one of the other and vice versa (Moreno-Hagelsieb and Latimer 2008). The RBBH compares well in term of accuracy with more complex approaches (Tekaia and Yeramian 2012; Altenhoff and Dessimoz 2009). Indeed, RBBH results are characterized by low rates of false positives. This method was therefore adopted here to find orthologs between the fully annotated genome of *A. vaga* and the partial sequences from the *A. ricciae* transcriptome.

We used the complete geneset of *A. vaga (*49,300 predicted genes) and 28,965 partial cDNA sequences obtained from hydrated and desiccated *A. ricciae* retrieved from Genbank. BLASTALL (TBLASTX with an E-value cutoff of 10^{-10}) was used as an alignment algorithm, yielding detection of 10,506 orthologs in total between *A. vaga* and *A. ricciae*. In order to detect orthologs of likely nonmetazoan origin, we used the Alien Index (AI) approach of Gladyshev and Meselson (2008) with the threshold set to 45 as in Flot et al. (2013). We found 955 orthologs with AI $\geq$ 45 and therefore considered them as putative HGT candidates shared by both *Adineta* species. The presence of 10 % shared foreign genes

between *A. ricciae* and *A. vaga* suggests that numerous HGTs happened already before these two lineages separated (unless some of them were acquired first by one lineage, then transferred horizontally to the other).

The genome of *A. vaga* is an ancient tetraploid in which 40 % of the genome is still organized in quartets of four homologous regions A1, A2, B1, and B2 (with A1-A2 and B1-B2 being allelic pairs and As being homologous to Bs; Flot et al. 2013). The nucleotide divergence in an allelic pair is low (average 3.8 %) whereas it is much higher in an ohnologous pair (average 26.4 %) (Flot et al. 2013). Interestingly, 47 orthologs with AI $\geq$ 45 were present in 4 copies, suggesting that HGT was already occurring before tetraploidy became established in bdelloids. When aligning (using BLASTX with a similarity threshold of 10^{-10}) these 47 orthologous genes against the transcriptome of the monogonont rotifer *B. plicatilis* (52,772 EST sequences available in GenBank; Denekamp et al. 2009; Suga et al. 2007), we found 3 hits (FM924334.1; FM931385.1; FM908274.1) suggesting that 3 genes had potentially been acquired by HGT in rotifers even before the separation of the bdelloid and monogonont clades. This hypothesis was confirmed when aligning these sequences against the whole GenBank database since the top five hits for each of them were bacteria or fungi (Table 10.2). The acquisition of foreign DNA by rotifers may therefore have started at the dawn of their evolutionary history.

After their integration in an eukaryotic genome, horizontally transferred genes of bacterial origin frequently acquire introns. Orthologs with AI $<$ 45 were characterized by a high proportion of introns: 93 % of them contained at least one intron and 57 % had more than three introns (Fig. 10.1). On the contrary, 32 % of the genes with AI $\geq$ 45 did not contain any intron and 26 % contained only 1 intron (Fig. 10.1). Out of the 47 orthologs with AI $\geq$ 45 present in 4 copies, only 7 did not contain any intron. These results suggest that intron acquisition is comparatively slower in bdelloids than in *M. incognita*, in which all putative HGT genes contained at least 1 intron.

10.6 Origins and Functions of Ancient Bdelloid HGTs: Emerging Results

A first functional analysis of the genes shared by *A. vaga* and *A. ricciae* was performed using the GOANNA pipeline (McCarthy et al. 2006). Since only partial sequences of *A. ricciae* were available, all subsequent analyses were done on the geneset of *A. vaga*. Briefly, orthologs were aligned using default parameters against the Swissprot database of manually curated protein sequences (comprising on 01/04/2014 542,258 sequences with 6,957,756 annotations). Each ortholog was attributed the gene ontologies (GO) of its top three hits in Swissprot. At the end of this process, 95 % of the orthologs with AI $<$ 45 and 86 % of the orthologs with AI $\geq$ 45 had received at least one GO term. To reduce the information content and gain a broader overview, each GO was replaced with its parent term in the PIR

Table 10.2 Top 5 best BLAST hits (using BLASTX) of 3 *B. plicatilis* EST homologous to *A. vaga/A. ricciae* sequences putatively acquired by HGT

Best BLASTX hits	E-value	%Identity	Taxonomic origin
FM931385.1			
Lactate dehydrogenase-like oxidoreductase (*Alistipes* sp. CAG:29)	2×10^{-46}	53	Bacteria
2-Hydroxyacid dehydrogenase (*Alistipes onderdonkii*)	2×10^{-46}	53	Bacteria
Lactate dehydrogenase-like oxidoreductase (*Alistipes finegoldii* CAG:68)	8×10^{-46}	51	Bacteria
2-Hydroxyacid dehydrogenase (*Bacteroidetes bacterium* oral taxon 272)	1×10^{-45}	53	Bacteria
2-Hydroxyacid dehydrogenase (*Bacteroides coprosuis*)	1×10^{-45}	51	Bacteria
FM908274.1			
Trp repressor binding protein (*Mucor circinelloides* f. *circinelloides* 1006PhL)	4×10^{-68}	58	Fungi
NAD(P)H:quinone oxidoreductase, type IV (*Mucor circinelloides* f. *circinelloides* 1006PhL)	7×10^{-68}	60	Fungi
Hypothetical protein RO3G_01460 (*Rhizopus delemar* RA 99-880)	5×10^{-66}	55	Fungi
NAD(P)H:quinone oxidoreductase, type IV (*Rhizopus delemar* RA 99–880)	3×10^{-65}	58	Fungi
NAD(P)H:quinone oxidoreductase, type IV (*Mucor circinelloides* f. *circinelloides* 1006PhL)	4×10^{-65}	56	Fungi
FM924334.1			
Sodium/potassium-transporting ATPase subunit alpha (*Mucor circinelloides* f. *circinelloides* 1006PhL)	2×10^{-53}	46	Fungi
Hypothetical protein MYCFIDRAFT_163458 (*Pseudocercospora fijiensis* CIRAD86)	2×10^{-49}	41	Fungi
Hypothetical protein PFL1_01803 (*Pseudozyma flocculosa* PF-1)	3×10^{-49}	44	Fungi
Hypothetical protein PPL_00295 (*Polysphondylium pallidum* PN500)	4×10^{-49}	44	Fungi
Hypothetical protein BC1G_04401 (*Botryotinia fuckeliana* B05.10)	1×10^{-48}	41	Fungi

slim ontology using GOSlimViewer from AgBase (McCarthy et al. 2006). The ten most abundant functions associated to orthologous genes with AI $\geq$ 45 were compared with their respective abundance in orthologous genes with AI $<$ 45 (Fig. 10.2). Interestingly, orthologs with AI $\geq$ 45 were enriched in putative oxidoreductase genes (+9 %) and hydrolases (+6 %). By contrast, protein-binding functions were underrepresented among orthologs with AI $\geq$ 45 compared with orthologs with AI $<$ 45 (15 % vs. 7 %). Protein-binding genes are involved in selective and noncovalent interactions with proteins or protein complexes (GO:0045308). These observations are in agreement with the previous observation that genes acquired horizontally and remaining functional in the new host are more often involved in specific enzymatic reactions than in regulatory processes and complex protein-protein interactions (Boschetti et al. 2012).

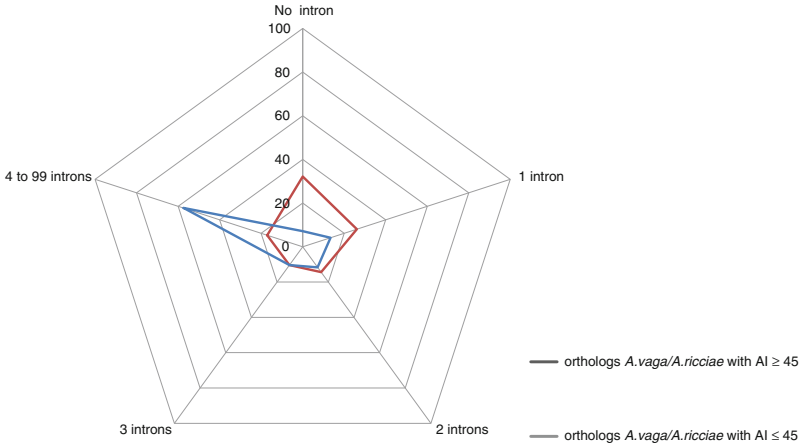


Fig. 10.1 *Distribution of the number of introns found in A. vaga/A. ricciae orthologs*. Introns were annotated in the *A. vaga* geneset by mapping RNAseq data (see Flot et al. 2013). Numbers represent percentages of orthologs

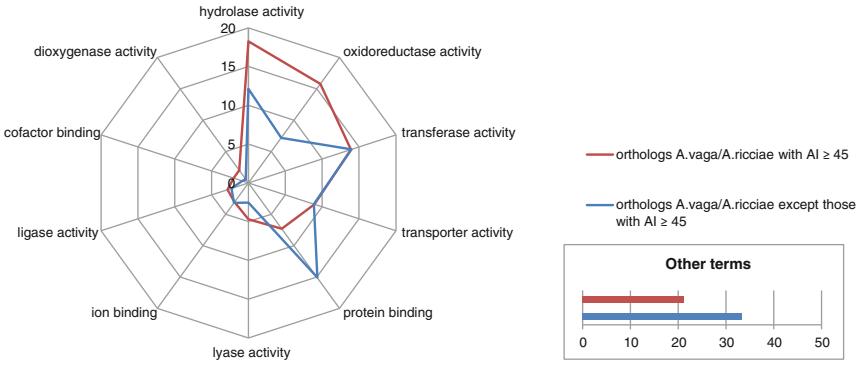


Fig. 10.2 *Top 10 most abundant GOSlim descriptions of the molecular functions of orthologs with AI ≥ 45 and with AI < 45*. Numbers represent percentages of orthologs. The inlet "other terms" represents the percentage of rare GOSlim terms not shown in the main graph

The abundance of hydrolase domains among genes of putative nonmetazoan origin can be explained by the high abundance of CAZymes (carbohydrate-related enzymes) in the genome of *A. vaga*, including a large number of glycoside hydrolases (Flot et al. 2013). At total of 299 orthologous CAZymes were found in both *A. vaga* and *A. ricciae*, including GH (glycoside hydrolases), GT (glycosyl transferases), PL (polysaccharide lyases), CE (carbohydrate esterases), and CBM (carbohydrate-binding modules). Of these 299 CAZymes shared between the two species, 99 appear to have been acquired through HGT (AI ≥ 45). Carbohydrates are a widely distributed source of energy and carbon. Therefore, the integration of

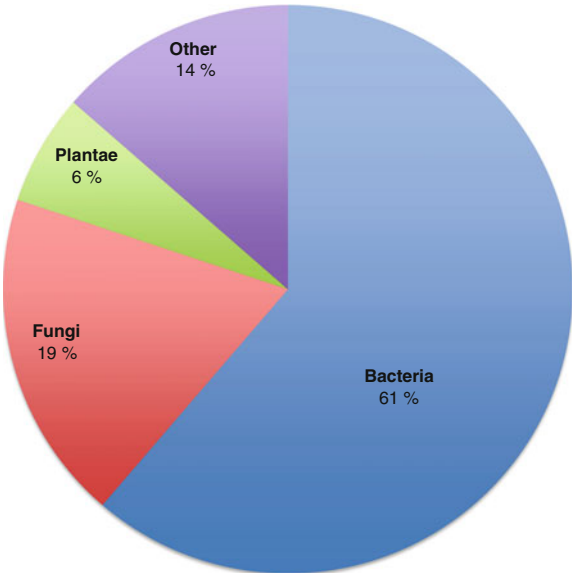


Fig. 10.3 *Pie chart of the putative donor groups of the horizontally acquired genes shared by A. ricciae and A. vaga*. Numbers represent percentages of orthologs. The most abundant donor group is bacteria, followed by fungi and plants. The donor group for each sequence was determined by reference to its best BLAST hit (smallest E-value) in GenBank

horizontally acquired CAZyme genes in bdelloids might have opened them access to new ecological niches by contributing to their adaptation to multiple food sources, as already observed in the case of human gut microbionts (Hehemann et al. 2012).

Oxidoreductases are involved in protection against oxidative damage. Several enzymes able to detoxify toxic free radical molecules were found among the *A. vaga/A. ricciae* orthologs, including nitric oxide dioxygenase, nitrilotriacetate monooxygenase, and nitroalkane oxidase. We screened orthologs for key antioxidant genes previously annotated in *A. vaga* (Flot et al. 2013): 30 % of these genes were found to have orthologs in the *A. ricciae* transcriptome. The 10 antioxidant gene families previously annotated were shared between the two *Adineta* species. In total, 19 % of the orthologs from three gene families were of putative non-metazoan origin: GST (glutathione S-transferase); GR (glutathione reductase) and AKR (aldo-keto reductase). By manual screening, we detected 6 orthologs with AI $\geq$ 45 that were putative homologs of trypanothione synthase, a key antioxidant in Kinetoplastida highlighted in Boschetti et al.'s transcriptome study. Several orthologs of three families of the antioxidant methionine sulfoxide reductase (MSR) were also found. As in *A. ricciae* (Boschetti et al. 2012), we were not able to detect any metazoan homolog of MSR in *A. vaga*. The metazoan homolog of another oxidoreductase acquired by HGT, stearoyl-CoA desaturase, was also absent in *A. ricciae* (Boschetti et al. 2012) and in *A. vaga*. This confirms the previous suggestion that HGTs can complement the absence of the corresponding metazoan gene (Boschetti et al. 2012). By contrast, screening the *B. plicatilis* transcriptome revealed the presence of a stearoyl-CoA desaturase expressed sequence (FM934536.1) of apparent metazoan origin, indicating that the

replacement of the metazoan stearoyl-CoA desaturase by a horizontally acquired homolog occurred in *Adineta* spp. after the divergence from monogononts but before the separation of the two species sequenced. In the case of MSR, each homologous sequence found in *B. plicatilis* was of apparent nonmetazoan origin as no metazoan hit was found in GenBank (using TBLASTX/BLASTX with a threshold of 10^{-10}). These observations support the hypothesis that the replacement of MSR by a non-metazoan homolog occurred in rotifers even before the separation of bdelloids and monogononts.

As a last analysis, a putative origin was assigned to each ortholog with AI $\geq$ 45 based on its best hit in GenBank. We found that 61 % were of apparent bacterial origin, 19 % of fungal origin, and 6 % from plants. Other orthologs were of diverse origins (Fig. 10.3). This predominantly bacterial origin of HGTs in rotifers had already been suggested in previous studies of *Adineta* (Boschetti et al. 2012; Flot et al. 2013).

10.7 Conclusions

In this chapter, the first comparison between the genes of two bdelloid species was performed to search for ancient acquisitions of nonmetazoan genes. The RBBH approach identified 10,506 orthologs between the partial transcriptome of *A. ricciae* and the predicted geneset of *A. vaga*. Among these orthologs, 9 % had an AI $\geq$ 45 and were therefore likely horizontal acquisitions. The presence of these genes in both species suggests that they were acquired prior to the divergence of *A. vaga* and *A. ricciae*. Furthermore, some of these genes of nonmetazoan origin are present in four copies in the genome of *A. vaga*, suggesting that HGTs were already happening before the tetraploidization event shared by all extant bdelloids. It is therefore highly probable that these genes will also be detected in other, more distantly related bdelloid species. Furthermore, we screened the transcriptomic data available for the monogonont *B. plicatilis* (a cyclical parthenogen) and found homologs to several bdelloid HGT candidates. This result supports the hypothesis that the acquisition of foreign genes predates the loss of sex in bdelloid rotifers. However, only 1.8 % of the genes of *B. plicatilis* seem to be of nonmetazoan origin (Boschetti et al. 2012): this is much lower than the 8–10 % reported in *A. vaga* and *A. ricciae*, suggesting that horizontal gene acquisitions are more frequent in rotifers that lack sex than in cyclical parthenogens (who do perform sex occasionally).

As mentioned previously, HGTs have been described in multiple metazoans and no consensus exists about how this foreign genetic material is acquired. Indeed, we speculate that these events are highly specific to the lifestyle of each organism. For example, most HGTs of *M. incognita* seem to have originated from bacteria living in the rhizosphere, in close proximity with this parasitic nematode (Castagnone-Sereno and Danchin 2014). Living in temporary habitats, bdelloids typically undergo multiple cycles of desiccation during their life as they are able to

withstand desiccation at any developmental stage. Adult monogononts, on the contrary, lack such ability but produce specific resting eggs during their sexual cycle that are protected against desiccation. Given that the genome of dehydrated bdelloids becomes shattered into multiple fragments that are subsequently repaired after rehydration (Hespeels et al. 2014), the frequent desiccation cycles experienced by bdelloids could facilitate their integration of foreign genes from the environment (Gladyshev and Meselson 2008; Hespeels et al. 2014). If such a mechanism at cellular level looks plausible, how extraneous DNA finds its way into the bdelloid germ line remains unclear. One possibility could be that foreign DNA present in the intestinal tract of bdelloids diffuses into the adjacent ovaries (the "you are what you eat" hypothesis; Doolittle 1998; Castagnone-Sereno and Danchin 2014) or perhaps even diffuses directly from the environment (Overballe-Petersen et al. 2013). Finally, how genes acquired from bacteria become associated with an eukaryotic promoter region (a prerequisite for expression) remains an intriguing question for which no satisfying hypothesis has been proposed yet.

As theoretically predicted and experimentally confirmed in our and previous studies, successful lateral gene transfers are more frequent for genes involved in metabolic processes than for those involved in DNA replication, transcription and translation (Boschetti et al. 2012; Whitaker et al. 2009; Jain et al. 1999). Indeed, a significant proportion of the shared HGTs of *A. vaga* and *A. ricciae* are involved in sugar metabolism, in antioxidant production and in detoxification, three processes essential to survive desiccation. The acquisition of these genes could therefore have been important events enhancing the desiccation resistance of bdelloids. This is in accordance with the hypothesis that HGT accelerates evolution and adaptation by extending the metabolic capabilities of the organisms, by improving their resistance to stress or by increasing their parasitic ability (Gogarten et al. 2002).

It has been hypothesized that, in the absence of meiosis and recombination, bdelloids and bacteria accumulate deleterious mutations. However, recent data from prokaryotes suggest that HGTs—known to occur frequently in bacteria and in bdelloids—can prevent Muller's ratchet from clicking by restoring genes inactivated by mutation and may therefore be an important prerequisite for the long-term maintenance of prokaryotic lineages (Takeuchi et al. 2014). A similar evolutionary role could be hypothesized for HGT in bdelloids. This is supported by the observation that the metazoan stearoyl-CoA desaturase has been replaced by its HGT equivalent in two *Adineta species*. (whereas the metazoan gene is still present in the monogonont *B. plicatilis*). Interestingly, the replacement of methionine sulfoxide reductase by its HGT homolog appears to have taken place in the common ancestor of bdelloids and monogononts prior to their divergence; however, additional phylogenetic studies are needed to confirm this hypothesis. Therefore, HGT could have a double role in bdelloids: first, speeding up evolution and adaptation, and second, counteracting Muller's rachet (with the help of gene conversion; cf. Flot et al. 2013).

References

Abad P et al (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat Biotechnol 26(8):909–915

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5(1):e1000262

Arkhipova IR, Meselson M (2005) Deleterious transposable elements and the extinction of asexuals. BioEssays 27(1):76–85

Arkhipova IR, Yushenova IA, Rodriguez F (2013) Endonuclease−containing *Penelope* retrotransposons in the bdelloid rotifer *Adineta vaga* exhibit unusual structural features and play a role in expansion of host gene families. Mob DNA 4(1):19

Birky CW Jr, (1996). Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. Genetics 144(1):427–437

Boschetti C et al (2012) Biochemical diversification through foreign gene expression in bdelloid rotifers PLoS Genet 8(11):e1003035

Castagnone-Sereno P, Danchin EGJ (2014) Parasitic success without sex—the nematode experience. J Evol Biol (in press)

Chapman JA et al (2010) The dynamic genome of *Hydra*. Nature 464(7288):592–596

Colbourne JK et al (2011) The ecoresponsive genome of *Daphnia pulex*. Science 331(6017):555–561

Connallon T, Clark AG (2010) Gene duplication, gene conversion and the evolution of the Y chromosome. Genetics 186(1):277–286

Crombach A, Hogeweg P (2007) Chromosome rearrangements and the evolution of genome structuring and adaptability. Mol Biol Evol 24(5):1130–1139

Danchin EGJ et al (2010) Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc Natl Acad Sci USA 107(41):17651–17656

Danchin EGJ et al (2011) Genomic perspectives on the long-term absence of sexual reproduction in animals. In: Pontarotti P (ed) Evolutionary biology—concepts, biodiversity, macroevolution and genome evolution. Springer, Heidelberg, pp 223–242

de Jonge R et al (2013) Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. Genome Res 23(8):1271–1282

Denekamp N et al (2009) Discovering genes associated with dormancy in the monogonont rotifer *Brachionus plicatilis*. BMC Genom 10(1):108

Doolittle FW (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14(8):307–311

Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78(2):737–756

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Biol 19(2):99–113

Flot, JF et al (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature 500(7463):453–457

Fontaneto D et al (2010) Cryptic diversity in the genus *Adineta* Hudson & Gosse, 1886 (Rotifera: Bdelloidea: Adinetidae): a DNA taxonomy approach. Hydrobiologia 662(1):27–33

Fontaneto D et al (2009) Extreme levels of hidden diversity in microscopic animals (Rotifera) revealed by DNA taxonomy. Mol Phylogenet Evol 53(1):182–189

Gladyshev EA, Arkhipova IR (2010) Genome structure of bdelloid rotifers: shaped by asexuality or desiccation? J Hered 101(Supplement 1):S85–S93

Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive horizontal gene transfer in bdelloid rotifers. Science 320(5880):1210–1213

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19(12):2226–2238

Gross SM, Williamson VM (2011) Tm1: a *Mutator/Foldback* transposable element family in root-knot nematodes. PLoS ONE 6(9):e24534

Haegeman A, Jones JT, Danchin EGJ (2011) Horizontal gene transfer in nematodes: a catalyst for plant parasitism? Molecular Plant-Microbe Interactions 24(8):879–887

Hehemann J-H et al (2012) Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. Proc Natl Acad Sci USA 109(48):19786–19791
Henry L, Schwander T, Crespi BJ (2012) Deleterious mutation accumulation in asexual *Timema* stick insects. Mol Biol Evol 29(1):401–408
Hespeels B et al (2014) Gateway to genetic exchange? DNA double-strand breaks in the bdelloid rotifer *Adineta vaga* submitted to desiccation. J Evol Biol (in press)
Hotopp JCD et al (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317(5845):1753–1756
Hsu WS (1956a) Oogenesis in *Habrotrocha tridens* (Milne). Biol Bull 111(3):364
Hsu WS (1956b) Oogenesis in the Bdelloidea rotifer *Philodina roseola* Ehrenberg. La Cellule 57:283–296
Hulsen T et al (2006) Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7(4):R31
Hur JH et al (2009) Degenerate tetraploidy was established before bdelloid rotifer families diverged. Mol Biol Evol 26(2):375–383
International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol 8(2):e1000313
Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96(7):3801–3806
Judson O, Normark B (1996) Ancient asexual scandals. Trends Ecol Evol 11(2):41–46
Khakhlova O, Bock R (2006) Elimination of deleterious mutations in plastid genomes by gene conversion. Plant J 46(1):85–94
Kikuchi T et al (2011) Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathog 7(9):e1002219
Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338
Kraaijeveld K et al (2012) Transposon proliferation in an asexual parasitoid. Mol Ecol 21(16):3898–3906
Leffler EM et al (2012) Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol 10(9):e1001388
Lesbarrères D (2011) Sex or no sex, reproduction is not the question. BioEssays 33(11):818
Lipinski KJ et al (2014) High spontaneous rate of gene duplication in *Caenorhabditis elegans*. Curr Biol 21(4):306–310
Lodé T (2011) Sex is not a solution for reproduction: the libertine bubble theory. BioEssays 33(6):419–422
Lunt DH (2008) Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins. BMC Evol Biol 8:194
Lynch M et al (2008) Localization of the genetic determinants of meiosis suppression in *Daphnia pulex*. Genetics 180(1):317–327
Maynard Smith J (1986) Evolution: contemplating life without sex. Nature 324(6095):300–301
Maynard Smith J (1978) The evolution of sex. Cambridge University Press, Cambridge
McCarthy FM et al (2006) AgBase: a functional genomics resource for agriculture. BMC Genom 7:229
Moran NA, Jarvik T (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328(5978):624–627
Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24(3):319–324
Muller HJ (1932) Some genetic aspects of sex. Am Nat 66(703):118–138
Ni T, Yue J, Sun G, Zou Y, Wen J, Huang J. (2012). Ancient gene transfer from algae to animals: Mechanisms and evolutionary significance. BMC Evol. Biol 12(1):83
Nikoh N et al (2010) Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet 6(2):e1000827
Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784):299–304

Overballe-Petersen S et al (2013) Bacterial natural transformation by highly fragmented and damaged DNA. Proc Natl Acad Sci USA 110(49):19860–19865
Paganini J et al (2012) Contribution of lateral gene transfers to the genome composition and parasitic ability of root-knot nematodes. PLoS ONE 7(11):e50875
Rödelsperger C, Sommer RJ (2011) Computational archaeology of the *Pristionchus pacificus* genome reveals evidence of horizontal gene transfers from insects. BMC Evol Biol 11(1):239
Schwander T, Crespi BJ (2009) Multiple direct transitions from sexual reproduction to apomictic parthenogenesis in *Timema* stick insects. Evolution 63(1):84–103
Schwander T, Henry L, Crespi BJ (2011) Molecular evidence for ancient asexuality in *Timema* stick insects. Current Biol 21(13):1129–1134
Segers H et al (2007) Annotated checklist of the rotifers (Phylum Rotifera), with notes on nomenclature, taxonomy and distribution. Zootaxa 1564:1–104
Segers H, Shiel RJ (2005) Tale of a sleeping beauty: a new and easily cultured model organism for experimental studies on bdelloid rotifers. Hydrobiologia 546(1):141–145
Seidl MF, Thomma BPHJ (2014) Sex or no sex: evolutionary adaptation occurs regardless. BioEssays 36(4):335–345
Stephens PJ et al (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144(1):27–40
Suga K et al (2007) Analysis of expressed sequence tags of the cyclically parthenogenetic rotifer *Brachionus plicatilis*. PLoS ONE 2(8):e671
Takeuchi N, Kaneko K, Koonin EV (2014) Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. G3 4(2):325–339
Tekaia F, Yeramian E (2012) SuperPartitions: detection and classification of orthologs. Gene 492(1):199–211
Triantaphyllou AC (1981) Oogenesis and the chromosomes of the parthenogenic root-knot nematode *Meloidogyne incognita*. J Nematol 13(2):95–104
Tucker AE et al (2013) Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. Proc Natl Acad Sci USA 110(39):15740–15745
Whitaker JW, McConkey GA, Westhead DR (2009) The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. Genome Biol 10(4):R36
Zhang J et al (2011) Transposable elements as catalysts for chromosome rearrangements. In Birchler JA (ed) Plant chromosome engineering. Methods in molecular biology. Humana Press, pp 315–326
Zhu B, Lou MM, Xie GL, Zhang GQ, Zhou XP, Li B, Jin GL (2011). Horizontal gene transfer in silkworm, *Bombyx mori*. BMC Genom 12(1):248

Part II
Phylogeography, Speciation and Coevolution

Chapter 11
Evolutionary History of Maternal Plant-Manipulation and Larval Feeding Behaviours in Attelabidae (Coleoptera; Curculionoidea) and Evolution of Plant-Basal Weevil Interaction

Chisato Kobayashi, Yudai Okuyama, Kazuhide Kawazoe, Masakado Kawata and Makoto Kato

Abstract Weevils are one of the dominant taxonomic groups in terrestrial ecosystem, diversifying to more than 60,000 described species. Although the most derived weevil group, Curculionidae, has adapted to and is utilising almost all parts of plant, basal groups show relatively limited larval feeding habits, such as pollen, seed, or fungus-infested wood feeders. Thus, it seems that ancestral larval infesting plant parts of weevils were restricted to N-rich, induced-defenseless, and temporal resources. Among the basal weevils, Attelabidae obtained such resources for their larvae by evolving unique maternal plant-manipulations: attelabid females manipulate specific young plant tissues of their host plants in a species-specific manner, e.g. cutting a shoot or a leaf, rolling a leaf, or constructing sophisticated wrapped leaf rolls, before and after oviposition presumably to secure

C. Kobayashi (✉) · M. Kawata
Division of Ecology and Evolutionary Biology, Graduate School of Life Sciences, The University of Tohoku, 3−6 Aoba, Aramaki, Aoba, Sendai, Miyagi 980-8578, Japan
e-mail: chisato.ck@gmail.com

M. Kawata
e-mail: kawata@m.tohoku.ac.jp

Y. Okuyama
Department of Botany, National Museum of Nature and Science, Amakubo 4−1−1, Tsukuba, Ibaraki 305-0005, Japan
e-mail: yokuyama@kahaku.go.jp

K. Kawazoe · M. Kato
Graduate School of Human and Environmental Studies, Kyoto University, Yoshida-Nihonmatsu-cho, Sakyo, Kyoto 606-8501, Japan
e-mail: kawazoe11@gmail.com

M. Kato
e-mail: kato@zoo.zool.kyoto-u.ac

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_11,

the survivorship of eggs or larvae. Molecular phylogenetic analyses based on the nuclear 18S and 28S ribosomal DNA and the mitochondrial COI genes indicated that the maternal plant-cutting behaviour originated in a common ancestor of Attelabidae, but was subsequently lost in the several lineages. The monophyly was recovered for the subfamily Attelabinae with high support, but not for the subfamily Rhynchitinae. By employing maximum likelihood-based ancestral state reconstructions, the larval leaf-blade feeding was inferred to have evolved from the boring of cut shoots/petioles. Moreover, the maternal leaf-rolling behaviour might have originated independently in the Attelabinae, Byctiscini, and also in several Deporaini lineages. Since the sophisticated behaviour of Attelabinae, i.e. constructing wrapped leaf rolls, have originated only once and not been lost in the lineage, these complex and innovative behaviours may have contributed to the success of the lineage diversification.

11.1 Evolution of Larval Feeding Habit in Basal Weevils

The superfamily Curculionoidea, i.e., weevils, is one of the most diverse insect groups, consisting of 62,000 described species, and, in addition, 158,000 more species estimated to exist (Oberprieler et al. 2007). Revealing the weevil-host plant relationships is an important challenge as this may contribute to our understanding of the evolution of herbivorous insects in relation to plant diversification (Morimoto 1962; Kuschel 1995; Wink et al. 1997; Marvaldi and Morrone 2000; Marvaldi et al. 2002; Hughes and Vogler 2004; Hundsdoerfer et al. 2009; McKenna et al. 2009).

Recent morphological and molecular works indicated that Curculionoidea contains seven families: Nemonychidae, Anthribidae, Belidae, Attelabidae, Caridae, Brentidae and Curculionidae (Oberprieler et al. 2007). The phylogenetic relationship of weevil families, based on McKenna et al. (2009), with number of described species of each family (Oberprieler et al. 2007) and the larval diet of each family are shown in Fig. 11.1. In the weevil families, the most basal group is Nemonychidae and Anthribidae, and these two families construct a sister group. Larvae of Nemonychidae are pollen feeders of conifers (Kuschel 1994; Oberprieler et al. 2007). Most larvae of Anthribidae feed on dying or dead fungus-infested branches and stems of angiosperms and gymnosperms, while some species feed on pollens, seeds, and even scale insects (Holloway 1982; Zimmerman 1994; Oberprieler et al. 2007). The family branched next to Nemonychidae and Anthribidae is Belidae. Larvae of Belidae are basically living, dying, or dead wood borers, or reproductive plant organ feeders (Zimmerman 1994; Oberprieler et al. 2007). Attelabidae is a unique group and larvae feed on dying young leaves, stems or seeds cut by mother insects. Larval feeding habits of Attelabidae are described in detail in the following section. Caridae is a relict family including only 18 species, and larvae feed on young seeds of conifers (Oberprieler et al. 2007). Brentidae is thought to have the ancestral angiosperm feeding, and most larvae feed on roots, living or dead woods, young stems or reproductive plant organs of angiosperms

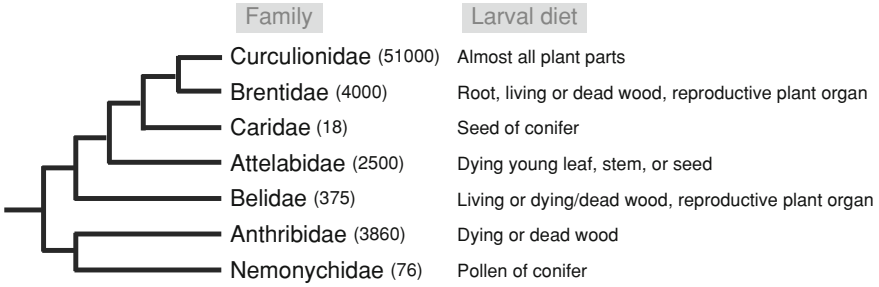


Fig. 11.1 Phylogeny of weevils and main larval diet of each family. Numbers in parentheses show the described species of each family

(Oberprieler et al. 2007). The most derived group, Curculionidae, show diverse larval feeding habit from virtually all plant-parts feeding to coprophagy or predation (Oberprieler et al. 2007).

The common trend in the larval diet of basal weevils is that they feed on nitrogen-rich resources such as pollens, seeds, or other reproductive plant organs, young leaves or stems, and fungus-infested plant parts. They are all high in nitrogen content (Mattson 1980; Slansky and Rodriguez 1987), but most of them are available only during the reproductive or foliation season of plants. Nitrogen content is important limiting factor for herbivorous insect, and the increase in nitrogen content enhance growth rate, reproduction, and survival rates (Mattson 1980; Awmack and Leather 2002). It seems that the basic demand for nitrogen-rich food in basal weevils might enable their rapid growth rate, high reproduction ability, and high survival rates, but their oviposition is limited by resource availability. Furthermore, there seems to be another trend in the plant parts consumed by larvae of basal weevils: they lack induced plant defense. When fresh plants, such as living leaves or stems, are damaged by herbivores, they increase their defensive compounds to avoid further damage (Karban and Baldwin 1997). Thus, feeding on living plants are required to conquer the induced plant defense. It seems that most basal weevils have no ability to deal with the induced plant defense. Considering that most derived weevil, Curculionidae, has adapted to almost all plant parts and exceptionally diversified, the evolutionary history of the larval feeding habit in weevils may be explained as the release from or being independent on nitrogen-rich and induced-defense-less resources.

11.2 Maternal Plant-Manipulation and Larval Feeding Behaviour in Attelabidae

Within weevils, the family Attelabidae is a unique group in which the females manipulate specific young structures of their host plants in various different ways before and after oviposition (Fig. 11.2). For example, attelabid females can cut a

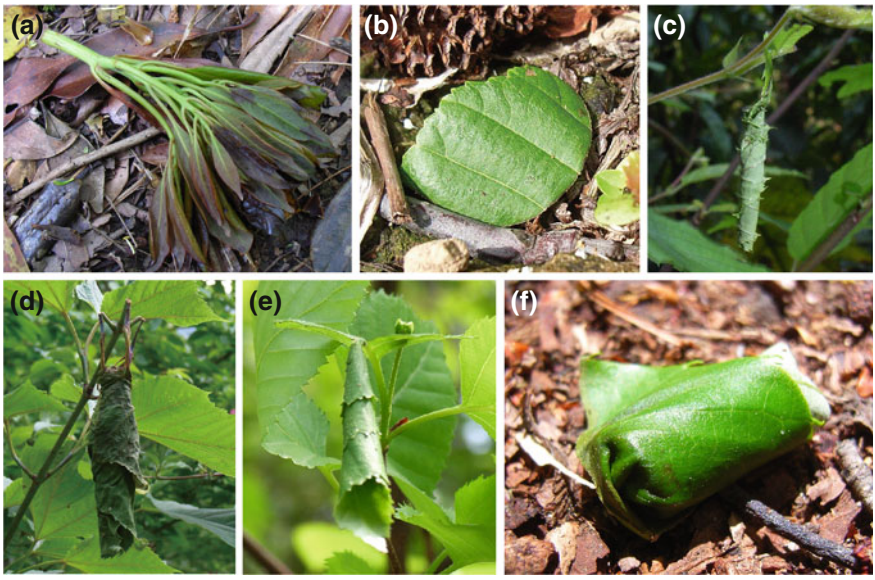


Fig. 11.2 Diverse plant-manipulation strategies of female attelabid weevils: **a** A shoot of *Machilus thunbergii* cut off by *Involvulus cornix*; **b** A leaf piece of *Alnus matsumurae* cut off by *Deporaus* sp. 2; **c** A leaf of *Carpinus laxiflora* cut, rolled up, and sealed by *Chonostropheus chujoi*; **d** Leaves of *Acer rufinerve* cut, rolled and glued together by a female *Bycticus venustus*; **e** A leaf of *Betula platyphylla* cut, rolled up, and sealed by *D. betulae*; **f** A tightly wrapped leaf roll constructed by a female *Henicolabus lewisii* using a leaf of *Quercus crispula*

bud/shoot/petiole/leaf, glue withered leaves together, or roll up a cut leaf blade (Kono 1930; Sakurai 1983; Vogt 1992; Hamilton 1994; Gønget 2003). These behaviours vary from simple shoot cutting to the sophisticated leaf-rolling strategies observed within the subfamily Attelabinae. Then, larvae grow only feeding on dying plant tissues cut by mother insects. Since larvae of basal weevils mainly feed on pollen and wood-decaying ascomycete fungi or woods (Nanninga 1991; Oberprieler et al. 2007), and larvae that feed on fresh plant tissue primarily occur in the family Curculionidae, which is the most derived family in Curculionoidea, as pointed out by Morimoto (1964), attelabid larval feeding on the dead plant tissue which is produced by maternal cutting seems to be an intermediate phase between dead-tissue feeders and fresh-tissue feeders. Constructing leaf rolls of Attelabinae requires highly complicated techniques, and the delicately wrapped leaf rolls of Attelabinae could be considered among the most skillful products constructed by animals (Fabre 1879−1907; Iwata 1935). The females measure and cut a leaf in a species-specific manner; they make many cuts on the midrib, pinch the leaf blade using their legs, fold the leaf lengthwise, and then roll up and fold the leaf into a cylindrical form without using silk or adhesives (Fig. 11.2f). In Bycticini, leaf rolls are constructed in different manners from Attelabinae: they cut petioles or stems using their long rostrum, and then roll and glue the withered leaves into a cigar-or

ball-like form (Fig. 11.2d). In Deporaini, although most species are leaf miners of cut leaves, some species also construct leaf rolls in a species-specific manner. For instance, species differ in the mode of measurement and leaf incision, and some species make holes by biting the leaf blade, and then roll the leaf into a funnel or cigar-like form (Fig. 11.2c, e). Although in other leaf rollers, such as lepidopteran larvae, leaf rolls are considered to reduce the risk of predation (Damman 1987) and solar radiation (Henson 1958), adaptive significance of the complicated leaf rolls constructed by Attelabidae remains unclear. Furthermore, leaf rolls of Attelabidae are different from those of other insects in that it is not a larva itself but a mother insect that constructs leaf rolls. The larvae of Attelabidae are generally endophytic herbivores, but their target host plant structures are diverse and can include buds, stems, petioles, leaf blades or seeds. Thus, the ecological and behavioural diversity and the remarkable leaf-rolling techniques make Attelabidae an interesting group for study. Revealing the evolutionary history of maternal plant-manipulation behaviours and larval feeding strategies may elucidate the underlying reasons why such sophisticated behaviours have evolved.

11.3 Molecular Phylogeny of Attelabidae

Phylogenetic relationships among attelabid species have been assessed based on morphological characters (Morimoto 1962, 1964; Sawada 1993; Legalov 2007). However, tribal tree topologies have differed among studies, and no phylogenetic tree of Attelabidae have been fully accepted (Oberprieler et al. 2007). To elucidate the evolutionary history of maternal plant-manipulation behaviours and larval feeding strategies in Asian fauna, we conducted a molecular phylogenetic analysis of 58 Japanese attelabid species, covering 75 % of the Asian 12 extant tribes and 60 % of the world's 15 extant tribes. The sampled attelabid weevil species, their localities, host plant species, plant-manipulation behaviours, target plant structures and sequenced regions are cited in Kobayashi et al. (2012). We followed Lawrence and Newton (1995) for family and subfamily classifications and Alonso-Zarazaga and Lyal (1999) for tribe classifications. Our samples included main groups that contain leaf-rolling species, i.e., Byctiscini, Deporaini and Attelabinae (Kono 1930; Legalov 2007). Thus, our sampling covered most of the plant-manipulation behavioural types observed within Attelabidae. Sequenced regions were a major portion (>75 %) of the nuclear 18S ribosomal DNA, as well as the D1 domain and flanking regions of 28S ribosomal DNA (18S and 28S, respectively) and mitochondrial cytochrome oxidase subunit I (COI) genes. Maximum parsimony (MP) trees were constructed from sequences for the three genes combined, and then, phylogenetic trees were estimated for the three genes combined using Bayesian methods. Detailed experimental and analysis methods are cited in Kobayashi et al. (2012).

The reconstructed tree (Fig. 11.3) strongly supported the monophyly of Attelabidae (84 % BS and 1.00 PP). Although the subfamily Rhynchitinae was not

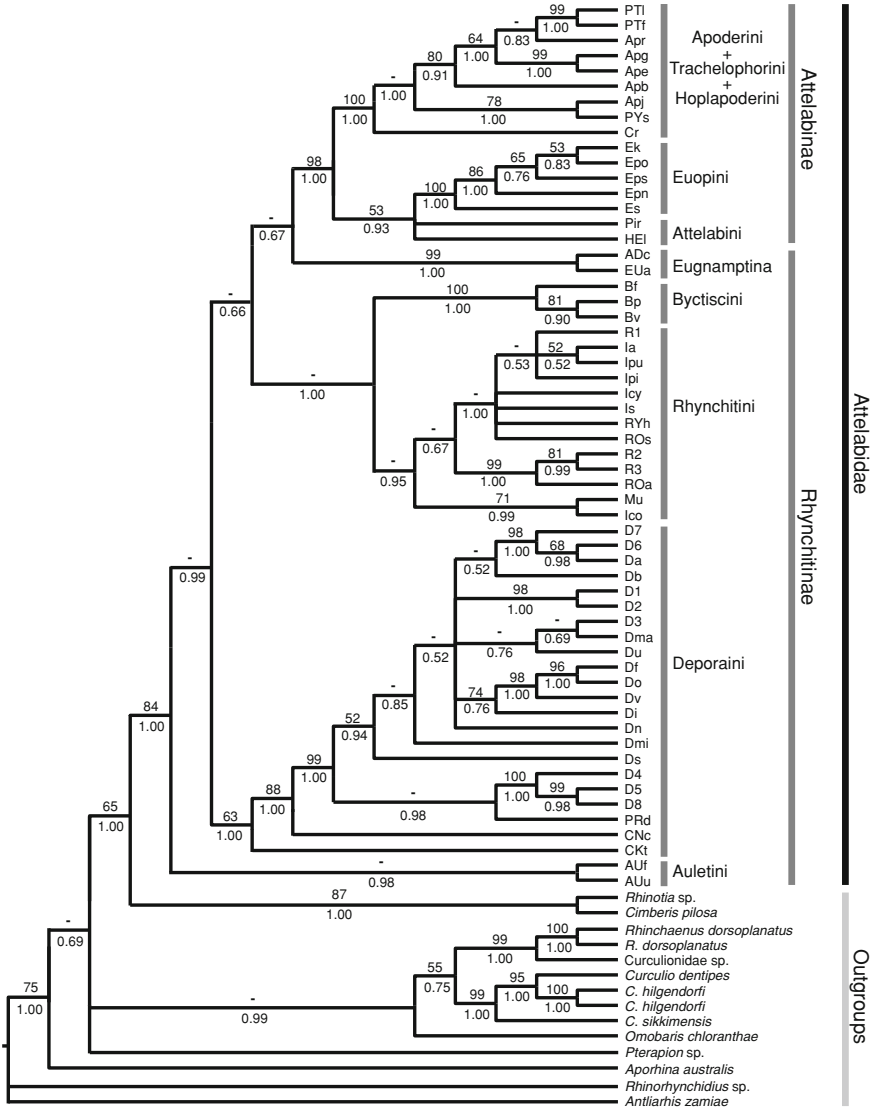


Fig. 11.3 A bayesian 50 % majority-rule consensus cladogram inferred from the nuclear 18S and 28S, and mitochondrial COI genes. Each species is represented by the corresponding code shown in Table 1 cited in Kobayashi et al. (2012). The numbers above branches indicate parsimony bootstrap values (displayed when >50 %), and those below branches indicate Bayesian posterior probabilities. *Vertical bars* to the right indicate taxa that correspond to tribes and subfamilies within Attelabidae. Reprinted from *Molecular Phylogenetics and Evolution*, vol. 64, Kobayashi et al. pp. 318−330. With permission from Elsevier

recovered as a monophyletic group, the subfamily Attelabinae was recovered as a monophyletic group with very strong support (98 % BS and 1.00 PP). The tribe Euopini and Byctiscini, and the subtribe Eugnamptina were strongly supported as monophyletic groups (≥99 % BS and 1.00 PP). The tribe Deporaini, with the exception of *Chokkirius truncatus* (Ckt), was recovered as a monophyletic group with strong support (88 % BS and 1.00 PP). Within the subfamily Attelabinae, the tribe Attelabini, Apoderini, Haplopoderini and Trachelophorini were not recovered as monophyletic. Within the subfamily Rhynchitinae, the tribe Rhynchitinae and Byctiscini were recovered as sister groups and the most basal split in Attelabidae leads to the tribe Auletini, although they were not supported in MP tree. Other tribe-level phylogenetic relationships within Rhynchitinae were ambiguous. The sister group of the subtribe Eugnamptina was also ambiguous. Although the oviposition behaviour of most Eugnamptina species remains unclear, some are reported to oviposit into dead or living leaves without any plant-manipulation (Hamilton 1980; Sawada 2000). Thus, Eugnamptina may be an ecologically unique group which lacks maternal plant-manipulation behaviour, and deserves to be classed as an independent tribe though its phylogenetic position remains unclear. Furthermore, with the exception of Euopini, the tribes within Attelabinae were not recovered as monophyletic. In order to gain more detailed insight into the phylogeny of Attelabidae, further phylogenetic analyses involving additional genes coupled with global taxon sampling is required.

11.4 Ancestral-Trait Reconstruction of Maternal Plant-Manipulation and Larval Feeding Behaviour

Maximum likelihood estimation of three ancestral traits, including maternal plant-manipulation and larval feeding strategies, was performed using BayesTraits (Pagel 1994, 1999; Pagel et al. 2004; available from www.evolution.rdg.ac.uk), following the procedures of Okuyama et al. (2008), by running the Maximum Likelihood analysis method with 100 interactions per tree. Detailed experimental and analysis methods are cited in Kobayashi et al. (2012). We used the Bayesian consensus tree for ancestral character mapping because within Attelabidae, 71 % (34 out of 48) of the nodes reconstructed in this study was strongly supported by PP value (>0.9). Maternal plant-manipulations were classified into two categories: cutting or not cutting the plant structure into which females oviposit. Further, maternal plant-manipulations were classified into another two categories: rolling or not rolling the leaf into/onto which females oviposit. Larval feeding structures were classified into two categories: leaf blades or other parts, such as the buds, petioles and stems.

11.4.1 Origin of Maternal Plant-Cutting Behaviour

The likelihood-ratio test for the model selection of the evolution of plant-cutting behaviour resulted in the selection of a one-parameter model with parameter restriction of q01 = q10 for the Bayesian consensus tree. The likelihood score obtained from the selected model was only 1.6 lower than the two-parameter model, indicating that there was no clear evolutionary trend in changes between cutting and not cutting behaviours. However, on all 21 nodes of the Bayesian consensus tree, the likelihood analysis strongly supported the presence of the maternal plant-cutting behaviour over absence of the behaviour (a difference of >2 units in log-likelihood scores). This suggests that the behaviour had a single origin at the common ancestor of Attelabidae and that it was subsequently lost from various lineages (Fig. 11.4). Considering that larvae of basal weevils basically rely on induced-defenseless resources, the plant-cutting behaviour of Attelabidae may have been innovative in the context of the nutritional and/or detoxification demands of larvae. Indeed, many herbivorous insects are known to cut their host plants and feed on the wilted parts to avoid plant defenses (Tallamy 1985; Becerra 1994; Dussourd and Denno 1994; Karban and Agrawal 2002; Becerra 2003). Thus, the acquirement of plant-cutting behaviour may have been contributed to the diversification of Attelabidae. Maternal plant-cutting behaviour would also play an important role by causing plant parts to wilt and soften: in leaf-rolling attelabine species, it has been observed that females wait to roll leaves after the cut plant parts have wilted (Iwata 1935). Furthermore, some attelabid weevils not only cut the plant tissues prior to oviposition, but also cut off the tissues after oviposition. As high levels of larval mortality have been observed in leaf miners and leaf rollers in the arboreal microhabitats due to parasitoid attacks (Kobayashi and Kato 2004a), maternal plant-cutting-off behaviour may have additionally evolved as a tactic to reduce parasitoid-induced larval mortality. This adaptiveness of larval feeding in plant parts, which were detached from the host plant, may explain the prevalence of maternal plant-cutting-off behaviour.

Due to the characteristic maternal plant cutting traits, attelabid larvae are regarded as decaying plant tissue feeders. Furthermore, some attelabids utilise fungi to accelerate or control the decaying process. In Attelabinae, females of Euopini store spores of symbiotic fungi in their mycangia and inoculate them on leaf rolls (Sakurai 1985; Sawada and Morimoto 1986; Riedel 2002; Kobayashi et al. 2008). Although the role of mycangial fungi remains unclear, the inferred monophyly of Euopini suggests that the relationship with symbiotic fungi may be traced back to the common ancestor of Euopini species.

The fact that very few species have lost the maternal plant-cutting behaviour suggests that the behaviour still remains necessary for attelabid larvae. However, some attelabids, including the Pterocolinae, Rhinchitini, Deporaini, and Eugnamptina groups, are known not to show maternal plant-manipulation (Hamilton 1980; Vogt 1992; Gønget 2003). Among the studied weevils, *Deporaus fuscipennis* (Df), *Paradeporaus depressus* (PRd), *Rhodocyrtus sanguinipennis* (ROs),

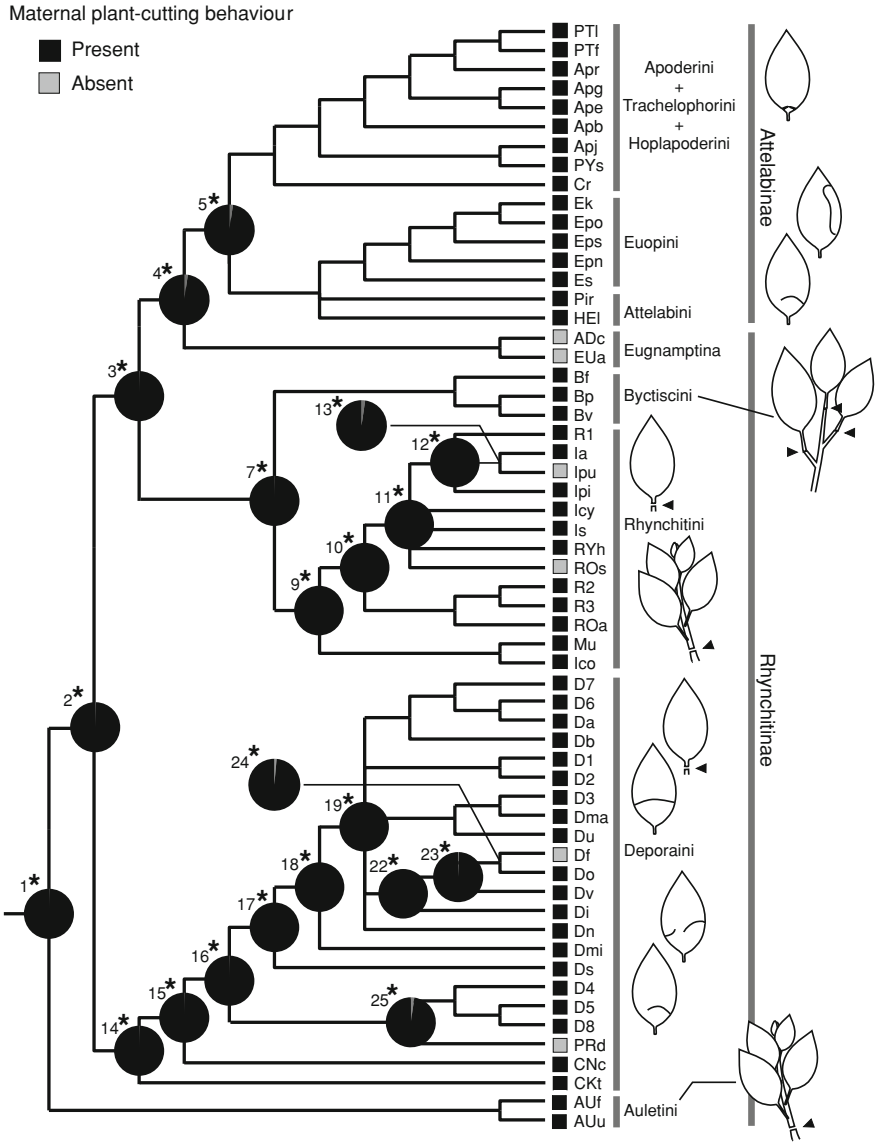


Fig. 11.4 Reconstruction of ancestral maternal plant-cutting behaviours. Pie charts illustrate the relative likelihoods (local estimators) of the 2 possible behavioural states of behaviours at each node, i.e. cutting plants (*black*) and not cutting plants (*grey*). Asterisks indicate strong support for the type over another type, judged from a difference of >2 units in log-likelihood scores. Illustrations on the *right-hand side* represent the typical pattern of maternal cutting and larval infestation site. Reprinted from *Molecular Phylogenetics and Evolution*, vol. 64, Kobayashi et al., pp. 318−330. With permission from Elsevier

Involvulus plumbens (Ipu), *Aderorhinus crioceroides* (ADc) and *Eugnamptus amurensis* (EUa) do not cut plant structures for oviposition. Reconstruction of ancestral traits based on the molecular phylogeny indicated these species have lost the maternal plant-cutting behaviour and such a loss of behaviour has independently occurred at least five times. Among the above species, ROs and Ipu are seed feeders, PRd is kleptoparasitic and exploits leaf rolls, which are constructed by Byctiscini, and ADc and EUa are estimated to be fallen-leaf miners or kleptoparasites of attelabine leaf rolls. The vascular bundle of plant tissues, which are fed upon by larvae of these species, have been already cut and are no more metabolically active, especially in terms of chemical defense, thus maternal plant-manipulation could be omitted. Df is the only species in which no host plant structures are manipulated, even though the larvae are leaf miners. The reasons why larvae of Df can develop in fresh leaves remain unclear. The fact that Df is abundant only in subalpine shrubs suggests that plant manipulation could be omitted in the extreme environments where levels of chemical defense are minimal due to scarcity of herbivores and low temperatures, and where leaf miners are almost free from parasitoid attack.

11.4.2 Evolution of Larval Infestation Sites

The likelihood-ratio test for the model selection of the evolution of larval infestation sites (leaf blades or other plant structures, $\alpha = 0.05$) resulted in the selection of a two-parameter model for the Bayesian consensus tree. Figure 11.5 show the relative likelihood between 2 categories of larval infestation sites on each of 15 focal nodes of the Bayesian consensus tree. The likelihood analysis strongly supported that larvae infested plant structures with the exception of leaf blades rather than leaf blades in the common ancestor of Attelabidae, Attelabinae + Eugnamptina + Byctiscini + Rhynchitini, and Deporaini (node 1, 3 and 14, respectively; a difference of >2 units in log-likelihood scores). On the other hand, the likelihood analysis strongly supported that larvae infested leaves over other plant structures in the common ancestor of Eugnamptina, Byctiscini and Deporaini except for CNc and CKt (nodes 6, 8 and 16, respectively; a difference of >2 units in log-likelihood scores). These results imply that there was clear evolutionary transition from feeders of other plant-parts, i.e. bud feeders, stem feeders, seed feeders, etc., into leaf-blade feeders. Since planate leaves are considered to have appeared and diversified at the end of the Palaeozoic period (Beerling et al. 2001) and foliar element-feeding was reported by fossil record in the Carboniferous (Labandeira 2002), when weevils had not yet appeared (McKenna et al. 2009), shifts of feeding sites into the leaf blades would have occurred following acquisition of the ability to utilise leaf blades rather than as an adaptation to a new niche. The ancestral larval feeding type of Attelabidae is inferred to be feeding of "manipulated" plant parts, except for leaf blades. Considering that larvae of Belidae, which is a sister group of Attelabidae and other derived weevils

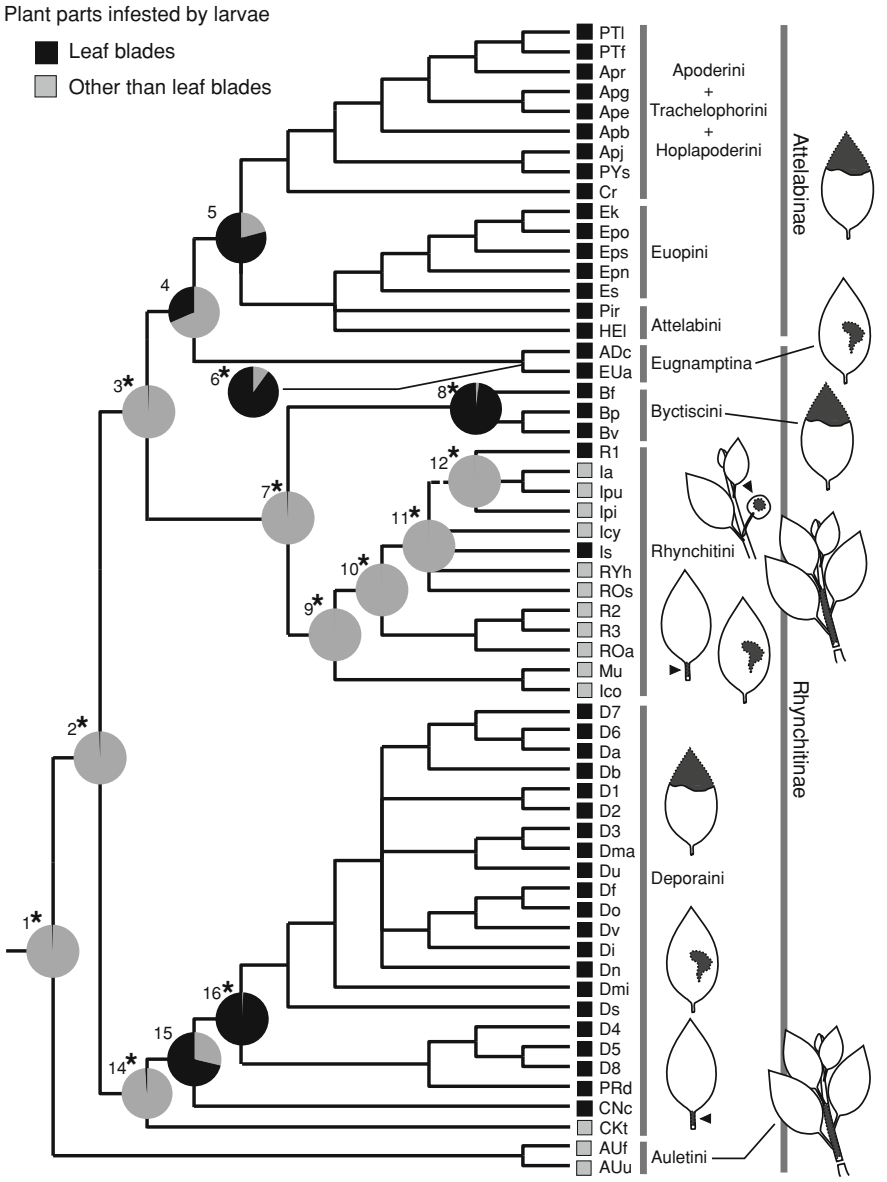


Fig. 11.5 Reconstruction of ancestral larval infestation plant parts. Pie charts illustrate the relative likelihoods (local estimators) of the 2 possible behavioural states of behaviours at each node, i.e. larval infestation of leaf blades (*black*) and plant-parts other than leaf blades (*grey*). Asterisks indicate strong support for the type over another type, judged from a difference of >2 units in log-likelihood scores. Illustrations on the *right-hand side* represent the typical pattern of maternal cutting. *Grey area* means the actual infestation area of each group/species. Reprinted from *Molecular Phylogenetics and Evolution*, vol. 64, Kobayashi et al., pp. 318−330. With permission from Elsevier

(Fig. 11.1), feed on stems or woods (Marvaldi et al. 2002, 2006), it is likely that cut shoot/petiole feeding is the most ancestral trait of attelabid larvae, and some lineages in Rhynchitini and Auletini lineages have retained the ancestral larval feeding trait. Generally, leaf is more desirable food for herbivores than shoot/petiole because of its softness and richer nutrition. In fact, mortalities other than predation and parasitism tend to be higher in shoot-feeding species than leaf-feeding species in Attelabidae (Kobayashi et al. unpublished) suggesting difficulties for early instar larvae to feed on the hard shoots. Thus, attelabid weevils may have evolved leaf-blade feeding as a result of adaptation to a plant part easier to digest. In gallwasps, it is reported that shifts of feeding sites within the same host plant promote speciation and diversification to a greater extent than host shifts do (Cook et al. 2002). Shifts of feeding sites into the leaf blades may also have been contributed to diversification of Attelabidae.

Although there is an evolutionary trend from shoot/petiole feeding towards leaf-blade feeding in Attelabidae, the direction is various among herbivorous insect groups. For example, in nematine sawflies, which induce galls on willows, there is an apparent evolutionary trend in galling site from the leaf edge towards the more central parts of the host plants like the petiole or stem (Nyman et al. 2000). In this case, it is suggested that galls evolved from leaf folders/rollers (Price 1992; Nyman et al. 2000). Further, in Lepidoptera, leaf mining is an ancestral trait and many linages secondarily evolved other plant-part feeding (Kristensen 1997; Connor and Taverner 1997).

11.4.3 Origin of Maternal Leaf-Rolling Behaviour

The likelihood-ratio test ($\alpha = 0.05$) for the model selection of the evolution of leaf-rolling behaviour resulted in the selection of a two-parameter model for the Bayesian consensus tree. Figure 11.6 shows the relative likelihood between two categories of presence versus absence of maternal leaf-rolling behaviour on each of 15 focal nodes of the Bayesian consensus tree. The likelihood analysis strongly supported that females did not roll leaves rather than that females rolled leaves in the common ancestor of Attelabidae, Attelabinae + Eugnamptina, Bycticini + Rhyncitini, and Deporaini (node 1, 4, 7, and 14, respectively; a difference of >2 units in log-likelihood scores). In contrast, the likelihood analysis strongly supported that females rolled leaves rather than did not roll leaves in the common ancestor of Bycticini (node 8, a difference of >2 units in log-likelihood scores). In the common ancestor of Attelabinae (node 5), although the likelihood that females rolled leaves was not strongly higher than the likelihood that females did not roll leaves, a higher likelihood that females rolled leaves was observed in1983 out of the last 2000 Bayesian trees. These results imply that the leaf-roll constructing behaviour originated independently more than 3 times in 3 clades: (1) a single origin of tightly wrapped leaf rolls in Attelabinae, (2) a single origin of sealed leaf rolls constructed from several tied leaves in Bycticini and (3) several origins of

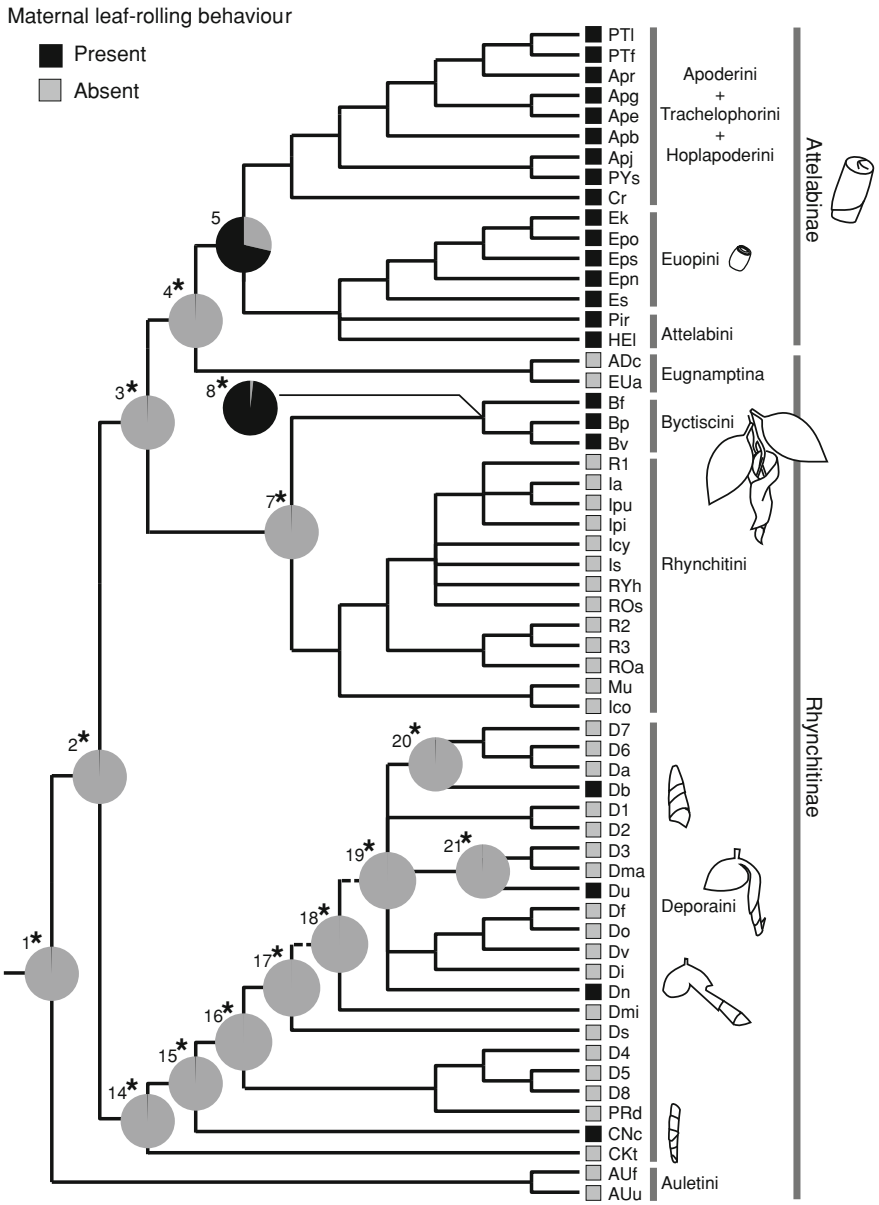


Fig. 11.6 Reconstruction of ancestral leaf-rolling behaviours onto the Bayesian majority-rule consensus tree for Attelabidae. Pie charts illustrate the relative likelihoods (local estimators) of the 2 possible behavioural states of behaviours at each node, i.e. constructing leaf rolls (*black*) and not-constructing leaf rolls (*grey*). Asterisks indicate strong support for the type over another type, judged from a difference of >2 units in log-likelihood scores. Illustrations on the *right-hand side* represent the typical pattern of leaf roll of each group/species. Reprinted from *Molecular Phylogenetics and Evolution*, vol. 64, Kobayashi et al., pp. 318−330. With permission from Elsevier

sealed or unsealed leaf rolls constructed by a single cut leaf in Deporaini. The former two leaf-rolling behaviours appear to have been conserved and have never been lost. These leaf rolls are constructed through a series of complex, time- and energy-consuming behaviours, e.g., measuring the size of the target leaf, cutting the leaf or petioles in a definite fashion, softening the leaf by nibbling it, folding the leaf longitudinally, rolling up the leaf by wrapping the leaf margin inward, or glueing withered leaves together. Furthermore, in these groups, species that construct leaf rolls in an intermediate phase are absent. This may be because intermediate leaf rolls are less adaptive than advanced leaf rolls and a lack of one behavioural step may result in the failure of construction of appropriate leaf rolls. This suggests that there may be a strong selection pressure to maintain a series of complex leaf-rolling behaviours. Considering the multiple origins of leaf-rolling behaviours, selection pressures to promote leaf-rolling behaviour should have existed.

It is intuitive to consider that leaf rollers may have evolved from species that developed leaf-blade feeding. In contrast to the predominantly exophytic larvae of the superfamily Chrysomeloidea, which is considered as the sister group of the superfamily Curculionoidea, curculionoid weevils evolved endophytic oviposition using their long rostrums and larval endophytic stages. For endophytic insects, leaves are not so safe due to their architectural thinness compared to other plant parts. In fact, leaf miners suffer from higher mortality due to parasitoid attacks relative to other more protected endophytic insects (Hawkins 1994). As the transition in infestation sites into leaf blades may have led to increased mortality as a result of increased exposure to natural enemies, especially parasitoids, leaf-rolling behaviour is considered to have evolved to facilitate larval survival such by adding outer shelters. Tightly wrapped and sealed leaf rolls presumably function to prevent predators and parasitoids from entering the leaf rolls. The interaction between attelabin species and their specific egg parasitoids may reflect an evolutionary arms race: females of these egg parasitoids evolved long ovipositor to penetrate walls of leaf rolls or can quickly slip into leaf rolls (Hirose 1968; Kobayashi and Kato 2004a, b). Additionally, leaf rolling has been reported to improve the nutritional quality of leaves for lepidopteran larvae (Sandberg and Berenbaum 1989; Sagers 1992). Thus, avoidance from natural enemies and/or improvement of food quality may explain the multiple origins of leaf-rolling behaviours in Attelabidae. Although the reason why the number of origin of leaf-rolling behaviour differ among linages remains unclear, species of Deporaini may have experienced different selection pressure from species of Attelabidae and Byctiscini.

Acknowledgments We wish to thank Dr. K Aoki, Dr. A Kawakita, Dr. H Nishi, Dr. R Goto and Y Kobayashi for supplying weevil samples. We also thank Dr. Y Kameda, and Dr. A Kawakita for their helpful advices regarding molecular techniques and the analysis of molecular phylogenies. We are grateful to Dr. H Kojima for identifying a curculionid weevil. We thank Dr. AA Legalov, Dr. A Riedel, Dr. Y Sawada and Dr. K Izawa for providing helpful and detailed information on weevil behaviours and lifestyles. This study was supported by JSPS Research Fellowships for Young Scientists.

References

Alonso-Zarazaga MA, Lyal CHC (1999) A world catalogue of families and genera of Curculionoidea (Insecta: Coleoptera). Entomopraxis, Barcelona

Awmack CS, Leather SR (2002) Host plant quality and fecundity in herbivorous insects. Annu Rev Entomol 47:817–844

Becerra JX (1994) Squirt-gun defense in Bursera and the chrysomelid counterploy. Ecology 75:1991–1996

Becerra JX (2003) Synchronous coadaptation in an ancient case of herbivory. Proc Natl Acad Sci USA 100:12804–12807

Beerling DJ, Osborne CP, Chaloner WG (2001) Evolution of leaf-form in land plants linked to atmospheric CO_2 decline in the late Palaeozoic era. Nature 410:352–354

Connor EF, Taverner MP (1997) The evolution and adaptive significance of the leaf-mining habit. Oikos 79:6–25

Cook JM, Rokas A, Pagel M, Stone GN (2002) Evolutionary shift between host oak sections and host-plant organs in Andricus gallwasps. Evolution 56:1821–1830

Damman H (1987) Leaf quality and enemy avoidance by the larvae of a pyralid moth. Ecology 68:88–97

Dussourd DE, Denno RF (1994) Host range of generalist caterpillars: trenching permits feeding on plants with secretory canals. Ecology 75:69–78

Fabre JH (1879–1907) Souveniers Entomologiques. 10:1823–1915

Gønget H (2003) The Nemonychidae, Anthribidae and Attelabidae (Coleoptera) of Northern Europe, Fauna Entomologica Scandivavica 38. Brill, Leiden Boston

Hamilton RW (1980) Notes on the biology of Eugnamptus collaris (Fabr.) (Coleoptera: Rhynchitidae), with descriptions of the larva and pupa. Coleopterists Bull 34:227–236

Hamilton RW (1994) New life cycle data for two western north American weevils (Coleoptera: Rhynchitidae), with a summary of north American Rhynchitid biology. Coleopterists Bull 48:331–343

Hawkins BA (1994) Pattern and process in host-parasitoid interactions. Cambridge University Press, Cambridge

Henson WR (1958) The effects of radiation on the habitat temperatures of some poplar-inhabiting insects. Can J Zool 36:463–478

Hirose Y (1968) Egg parasitism of some leaf-rolling weevils in relation to the number of eggs laid in a leaf roll, with special reference to parasitism by Poropoea morimotoi Hirose (Hymenoptera: Trichogrammatidae). Kontyu 36:377–388 (in Japanese)

Holloway BA (1982) Anthribidae (Insecta: Coleoptera). Fauna of New Zealand 3:1–264

Hughes J, Vogler AP (2004) The phylogeny of acorn weevils (genus Curculio) from mitochondrial and nuclear DNA sequences: the problem of incomplete data. Mol Phyl Evol 32:601–615

Hundsdoerfer AK, Rheinheimer J, Wink M (2009) Towards the phylogeny of the Curculionoidea (Coleoptera): reconstructions from mitochondrial and nuclear ribosomal DNA sequences. Zool Anz 248:9–31

Iwata K (1935) On the habits of some Rhynchitinae, Attelabinae and Apoderinae in Japan. Kontyu 9:261–278 (in Japanese)

Karban R, Agrawal AA (2002) Herbivore offense. Annu Rev Ecol Syst 33:641–664

Karban R, Baldwin IT (1997) Induced responses to herbivory. University of Chicago Press, Chicago

Kobayashi C, Kato M (2004a) To be suspended or to be cut off? Differences in the performance of two types of leaf-rolls constructed by the attelabid beetle Cycnotrachelus roelofsi. Popul Ecol 46:193–202

Kobayashi C, Kato M (2004b) A new species of Poropoea (Trichogrammatidae) oviposites by entering through the oviposition hole of attelabid beetle. Contr Biol Lab Kyoto Univ 29:431–436

Kobayashi C, Fukasawa Y, Hirose D, Kato M (2008) Contribution of symbiotic fungi to larval nutrition of a leaf-rolling weevil. Evol Ecol 22:711–722

Kobayashi C, Okuyama Y, Kawazoe K, Kato M (2012) The evolutionary history of maternal plant-manipulation and larval feeding behaviours in attelabid weevils (Coleoptera; Curculionidae). Mol Phyl Evol 64:318–330

Kono H (1930) Die biologischen gruppen der Rhynchitinen, Attelabinen und Apoderinen. J Fac Agr Hokkaido Univ 29:1–36

Kristensen NP (1997) Early evolution of the Lepidoptera + Trichoptera lineage: phylogeny and the ecological scenario. In: Grandcolas P (ed) The origin of biodiversity in Insects: phylogenetic tests of evolutionary scenarios. Mémoires du Muséum National d'Histoire Naturelle 173, pp 253–271

Kuschel G (1994) Nemonychidae of Australia, New Guinea, and New Caledonia. In: Zimmerman EC (ed) Australian weevils. Volume I—Orthoceri, Antribidae, to Attelabidae. Primitive weevils. CSIRO, Australia, pp 563–637

Kuschel G (1995) A phylogenetic classification of the Curculionoidea to families and subfamilies. Mem Entomol Soc Washington 14:5–33

Labandeira CC (2002) The history of associations between plants and animals. In: Herrera CM, Pellmyr O (eds) Plant-animal ineteractions: an evolutionary approach. Wiley-Blackwell, pp 26−74

Lawrence JF, Newton AF Jr (1995) Families and subfamilies of Coleoptera (with selected genera, notes, references and data on family-group names). In: Pakaluk J, Slipinski SA (eds) Biology, phylogeny, and classification of coleoptera: papers celebrating the 80th birthday of Roy A. Crowson, vol. 2. Muzeum I Instytut Zoologii PAN, Warszawa, pp 779−1006

Legalov AA (2007) Leaf-rolling weevils (Coleoptera: Rhynchitidae, Attelabidae) of the world fauna. Novosibirsk

Marvaldi AE, Oberprieler RG, Lyal CHC, Bradbury T, Anderson RS (2006) Phylogeny of the Oxycoryninae s.l. (Coleptera Phytophaga) and evolution of plant-weevil interactions. Invertebr Syst 20:447–476

Marvaldi AE, Morrone JJ (2000) Phylogenetic systematics of weevils (Coleoptera: Curculionoidea): a reappraisal based on larval and adult morphology. Insect Syst Evol 31:43–58

Marvaldi AE, Sequera AS, O'Brien CW, Farrell BD (2002) Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): Do niche shifts accompany diversification? Syst Biol 51:761–785

Mattson WJ (1980) Herbivory in relation to plant nitrogen content. Ann Rev Ecol Syst 11:119–161

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. Proc Natl Acad Sci USA 106:7083–7088

Morimoto K (1962) Comparative morphorogy and phylogeny of the superfamily Curculionoidea of Japan. J Fac Agr Kyusyu Univ 11:331–373

Morimoto K (1964) Characteristics and evolution of oviposition behavior of Attelabidae. Insect J 1:15–21 (in Japanese)

Nanninga F (1991) Superfamily curuculionoidea. In: Csiro (eds) The insects of Australia vol 2, 2nd ed. Melbourne University Press, Australia, pp 678–683

Nyman T, Widmer A, Roininen H (2000) Evolution of gall morphology and host-plant relationships in willow-feeding sawflies (Hymenoptera: Tenthredinidae). Evolution 54:526–533

Oberprieler RG, Marvaldi AE, Anderson RS (2007) Weevils, weevils, weevils everywhere. Zootaxa 1668:491–520

Okuyama Y, Pellmyr O, Kato M (2008) Parallel floral adaptations to pollination by fungus gnats within the genus Mitella (Saxifragaceae). Mol Phyl Evol 46:560–575

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc B 225:37–45

Pagel M (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst Biol 48:612–622

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673–684
Price PW (1992) Evolution and ecology of gall-inducing sawflies. In: Shorthouse JD, Rohfritsch O (eds) Biology of insect-induced galls. Oxford University Press, Oxford, pp 208–224
Riedel A (2002) Taxonomy, phylogeny, and zoogeography of the weevil genus Euops (Insecta: Coleoptera: Curculionoidea) in the Papuan region. PhD Thesis, Ludwig Maximilians University, Munich
Sagers CL (1992) Manipulation of host plant quality: herbivores keep leaves in the dark. Funct Ecol 6:741–743
Sakurai K (1983) Ethological and ecological studies of leaf-roll weevils and other small animals. Kyoto Univ, Ethol Res of Small Animals in the Humid Tropics, pp 129–146
Sakurai K (1985) An attelabid weevil (*Euops splendida*) cultivates fungi. J Ethol 3:151–156
Sandberg SL, Berenbaum MR (1989) Leaf-tying by tortricid larvae as an adaptation for feeding on phototoxic Hypericum perforatum. J Chem Ecol 15:875–885
Sawada Y (1993) A systematic study of the family Rhynchitinae of Japan (Coleoptera, Curculionoidea). Hum Nat 2:1–93
Sawada Y (2000) Ecological notes on Eugnamptus flavipes (Sharp, 1889) (Coleoptera: Rhynchitidae), with description of the larva. Nat Hum Activities 5:1–3
Sawada Y, Morimoto K (1986) The mycetangia and the mode of the fungus transmission in the weevil genus Euops (Coleoptera: Attelabidae). Sci Bull Fac Agr Kyushu Univ 40:197–205 (In Japanese)
Slansky F Jr, Rodriguez JG (1987) Nutritional ecology of insects, mites, spiders, and related invertebrates. Wiley, New York, pp 449–486
Tallamy DW (1985) Squash beetle feeding behavior: an adaptation against induced cucurbit defenses. Ecology 66:1574–1579
Vogt GB (1992) Leaf-rolling weevils (Coleoptera: Attelabidae), their host plants, and associated rhynchitid weevils in North America (Canada through the Republic of Panama): summary of a long-term field study. In: Quintero D, Aiello A (eds) Insects of panama and mesoamerica. Oxford University Press, New York, pp 392–420
Wink M, Mikes Z, Rheinheimer J (1997) Phylogenetic relationships in weevils (Coleoptera: Curculionoidea) inferred from nucleotide sequences of mitochondrial 16S rDNA. Naturwissenschaften 84:318–321
Zimmerman EC (1994) Australian weevils. Volume I—Orthoceri, Antribidae, to Attelabidae. Primitive weevils. CSIRO, Australia

Chapter 12
Microevolution of Insect–Bacterial Mutualists: A Population Genomics Perspective

Amanda M. V. Brown

Abstract The advent of high-throughput sequencing has ushered in a new era with enormous promise for study of short-timescale evolution. Insect–bacterial mutualists typically have tiny genomes allowing deep coverage for large populations. The importance of mutualists in ecological phenomena, such as invasions, leave much room for study. Nutritional mutualists of Order Hemiptera are of interest for being long coevolved with their hosts, thus degenerated in gene content, yet able to support their hosts on varying diets and during invasions. Determining the symbiont's role and how adaptation might be achieved with a limited genetic repertoire is a challenge. One fruitful approach is highlighted in a recent population genomic study using a natural experiment: the 2009—present spread of the U.S. soybean pest *Megacopta cribraria* (Hemiptera: Plataspidae). This invasive species' symbiont *Ishikawaella capsulata* (Gamma-proteobacteria) was previously shown to determine the pest-status of its host. Deep sequencing revealed allele frequency change since arrival that matched predictions, showing signatures of purifying and positive selection, with differences in "symbiont role" genes associated with different host plants. In the near future, this approach applied to other systems may illuminate the important role and dynamic potential of microbial symbionts in ecosystems.

12.1 General Background and Introduction

Mutualism, defined as a relationship in which two unrelated species mutually benefit from their interaction, is clearly widespread and important in nature (Aslan et al. 2013; Ferrari and Vavre 2011; Goodrich-Blair and Hussa 2013) and can take

A. M. V. Brown (✉)
Department of Integrative Biology, Oregon State University (OSU), 3029 Cordley Hall, Corvallis, OR 97331, USA
e-mail: amvbrown@gmail.com

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_12,

on an enormous variety of forms. Intimate mutualisms in which organisms live inside or on others throughout their life cycles are most easily recognized and studied, whereas less intimate associations may be missed (Herre et al. 1999). Microbial mutualisms are also especially likely to be missed (Orphan 2009). Undoubtedly, there are cryptic mutualisms whose role may be essential to the host and perhaps even the ecosystem (Palmer et al. 2010). Despite the ubiquity of mutualism, we are at the early stages of measuring the functional roles and evolutionary dynamics of mutualists using genomics tools.

12.2 Introduction to the Sap-Feeding Insects and Their Mutualists

A number of recent reviews have highlighted the insects in the Order Hemiptera (e.g., aphids, true bugs, cicadas, etc.) for their exceptionally coevolved mutualists (Dale and Moran 2006; MacDonald et al. 2011; McCutcheon and Moran 2010). These insects are constrained by their piercing and sucking mouthparts to liquid diets, usually consisting of plant sap or sometimes blood. Since these food sources are imbalanced, lacking essential amino acids and vitamins, the Hemiptera have developed long-evolved partnerships with nutritional mutualist bacteria that synthesize the missing nutrients in their diets (Baumann 2005; Buchner 1965). The extremely long duration of these associations has led to such extreme degradation in genome and cellular features that some really have more in common with organelles than free-living organisms, e.g., *Nasuia deltocephalinicola* from the leafhopper *Macrosteles* sp. with its 0.112 Mb genome, the smallest known for a cellular organism (Bennett and Moran 2013).

There are some puzzling questions in this group, in particular, how can these highly degenerate symbiont genomes adjust nutrient supply to the insect host as needed for those with generalist diets? Can the symbionts adapt to changes in the host diet? The answers are not clear (Hansen and Moran 2013), but the prediction is that the vertical transmission, lack of recombination and genome reduction leave a limited genetic repertoire for adaptation (Andersson 2008; McCutcheon and Moran 2011; Mira et al. 2001). Thus, how is it possible that so many hemipterans (e.g., whiteflies, aphids, scale insects, psyllids, stinkbugs, bed bugs, etc.) are such successful invasives (Pimentel et al. 2005)? Part of the explanation for the success of these pests could be the presence of additional facultative symbionts that supplement the limited genetic repertoire of the obligate nutritional symbiont (Russell et al. 2013). Another explanation could be the reliance on dual symbioses in many of these species (McCutcheon and Moran 2010).

Searching for potential adaptive capacity in symbionts of sap-feeding insects slightly may be easier beginning with those having less-degraded genomes. For example, not beginning with the tiny *Nasuia* from leafhoppers or *Hodgkinia* from cicadas (genome sizes 0.112 and 0.14 Mb), but instead beginning with symbionts

with reasonably "large" genomes (~0.6–0.8 Mb) in *Buchnera* from aphids and *Ishikawaella* from the kudzu bug.

12.2.1 What Is the Genomic Signature of a Microbial Mutualist?

Genome reduction is universal in the obligate mutualists of the sap-feeding insects, both intracellular and extracellular (Moran et al. 2008; Nikoh et al. 2011). It is presumed that the main forces contributing to this are (1) reduced effective population size reducing the opportunity for genetic exchange and thereby leading to drift and a ratchet effect (Mira and Moran 2002) and (2) a universal bias toward deletions in bacteria (Mira et al. 2001). The resulting genome effects are continuous pseudogenization and losses of coding and regulatory sequences for all but the most essential genes. It might be assumed that the genome size distribution in extant hemipteran obligate symbionts is correlated with the age of the association within the host, but this has not been explicitly shown.

One genomic signature generally missing in heritable endosymbionts, quite in contrast to organelles, which also originated as endosymbionts, is massive horizontal gene transfer to the host nucleus, although some horizontal gene transfer does occur (Dunning Hotopp 2011; Husnik et al. 2013; Nikoh et al. 2010).

12.2.2 Adaptation with Limited Genetic Repertoire?

Clearly, the limited genetic content in obligate microbial symbionts could be a challenge when hosts encounter new habitats. Indeed, microbial symbionts often accompany invasive species (Feldhaar 2011; Janson et al. 2008; Lichman 2010; Richardson et al. 2000; Travaset and Richardson 2011); however, few studies examine how genomes evolve in these symbionts during invasions (Bennett 2013; Desprez-Loustau et al. 2007). It is unclear to what extent such symbionts could drive the evolutionary trajectory of the partnership (Clay and Holah 1999), but clearly this question warrants further study, particularly in the case of invasives.

12.2.3 Population Genomics of Microbial Mutualists: What Predictions Can We Make?

Population genetic studies of microbial mutualists using small numbers of markers have generally highlighted the lack of adaptability of obligate symbionts compared with facultative symbionts (Abbot and Moran 2002; Funk et al. 2001).

If full-genome studies reveal more genetic variation, what is the theoretical backdrop for population dynamic predictions? First, nutritional symbionts that provide a limiting resource should be subject to strong purifying selection on these essential nutrient pathways, even in the face of large drift. Also, strong bottleneck and resulting drift may facilitate faster accumulation of beneficial mutations, even in genes undergoing high-average levels of purifying selection.

Lastly, because selection on microbial symbionts happens at multiple levels (e.g., at the inter-symbiont level and at the inter-host level), population genomics predictions may be divided into those relating to genes for microbe survival in a host and those relating to symbiont role for host survival in its environment. Because of this inherent complexity, the study of population genomics of microbial mutualists may initially benefit from examining simplified systems—for example, those with just one primary symbiont and perhaps just two host environments. The following study (Brown et al. 2013) is an example of this.

12.3 A Natural Experiment: The Invasive Pest *Megacopta cribraria* and its Microbial Mutualist *Ishikawaella capsulata*

This section reviews a recent study of an obligate microbial mutualist that follows from the discussion above: a natural experiment (invasion) involving an insect host with just one nutritional symbiont and two host plants. This natural experiment allowed researchers to examine adaptive potential of a symbiont with a limited genetic repertoire. What makes this study system especially suitable for the questions above is that its symbiont exerts a clear influence on the host, causing a change in its pest-status.

The plataspid stinkbug *M. cribraria* (Hemiptera: Plataspidae), now known as the kudzu bug in the U.S., is a phloem-feeding legume (Fabaceae) specialist. It is endemic to Europe and Asia and was only first discovered in the New World in Hall County, Georgia, USA in October 2009 (Eger et al. 2010). Initially, this invasive insect spread rapidly in Georgia on its native wild plant host, kudzu (*Pueraria* spp.), which itself is an introduced invasive plant (Forseth and Innis 2004). However, within a year, it spread to cultivated soybeans (*Glycine max*), becoming a significant pest (Kikuchi and Kobayashi 2010; Suiter et al. 2010; Zhang et al. 2012). It causes up to 32.8 % reduction in kudzu growth and 19 % decrease in soybean yield, a crop with ~$40 billion value in the U.S. (Zhang et al. 2012; United States Department of Agriculture 2012). In just 4 years, from 2009 to 2013, *M. cribraria* has spread to 12 U.S. States, infesting an area of about 890,000 km^2 (University of Georgia—Center for Invasive Species and Ecosystem Health 2013) (see Fig. 12.1). Insect phenotype varies, sometimes apparently according to plant (see Fig. 12.2).

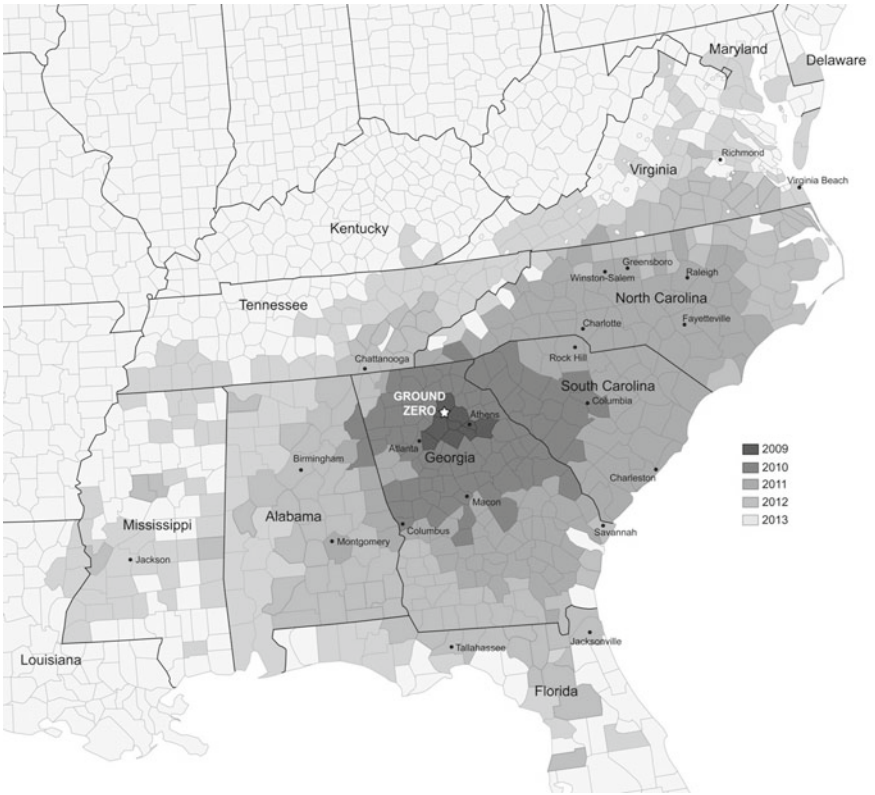


Fig. 12.1 Map of 2009–2013 spread of *M. cribraria* in the U.S. by county and state, showing the original discovery location, marked "ground zero"

M. cribraria, like most hemipterans, has a diet constrained by its piercing and sucking mouthparts: it can feed only on plant phloem. The nutrient profile of phloem is generally severely depleted in essential amino acids and vitamins (Sandström and Pettersson 1994). Hence, this insect depends on its coevolved obligate gut-symbiont *Candidatus I. capsulata* (Gamma-proteobacteria) for survival (Fukatsu and Hosokawa 2002). The symbiont *Ishikawaella* presumably synthesizes missing nutrients in the insect's diet, although determination of the exact contribution by the symbiont has not yet been demonstrated.

12.3.1 A Mutualist Whose Genotype Determines its Host Phenotype

The most remarkable feature of this system is that experiments have shown the mutualist determines its host's phenotype. In Japan, *M. cribraria* is found in the

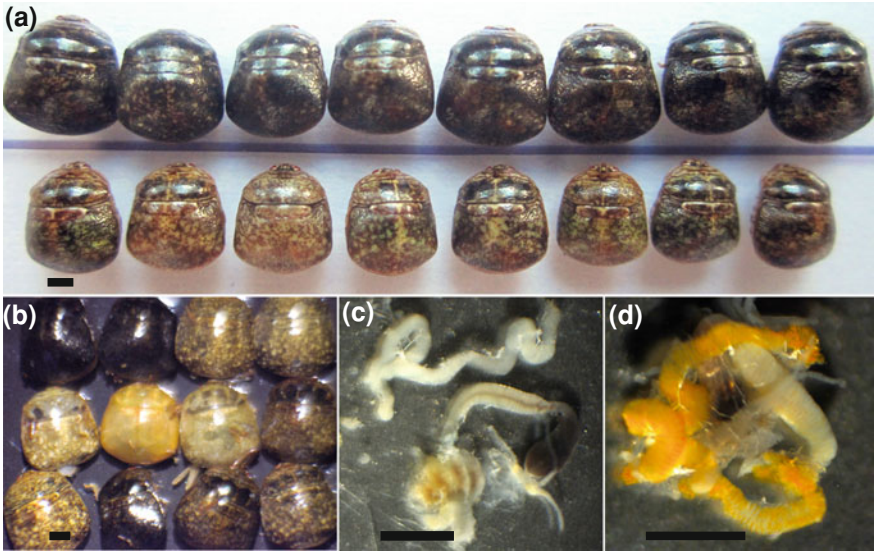


Fig. 12.2 *M. cribraria* in the U.S. showing phenotype variation. **a** Females on soybeans *top row*, females on nearby kudzu, *bottom row*. **b** Adults on a single patch of kudzu. **c**, **d** Specialized midgut section with symbiont-bearing crypts, showing color range from whitish to vibrant orange, the latter being commonest on soybeans

southern islands and is rarely a pest, whereas the closely related species *M. punctatissima* in the northern islands is a frequent soybean pest (Hosokawa et al. 2007). Experimental swap of the symbionts between the nonpest *M. cribraria* and the pest *M. punctatissima* demonstrated a reversal of the phenotype, whereby the pest insect showed reduced hatch rate on soybeans with the nonpest symbiont, and conversely, the nonpest insect had increased hatch rate on soybean with the symbiont from the pest (Hosokawa et al. 2007). It is the life history of the symbiont *Ishikawaella* that made it possible to demonstrate this experimentally: it lives extracellularly in bacteria-filled crypts in a special section of the gut (Fukatsu and Hosokawa 2002). Females transmit *Ishikawaella* vertically to offspring by laying symbiont-filled "capsules" along with eggs. Newly hatched nymphs must feed on the capsules and ingest an inoculum of symbionts to last until adulthood (Hosokawa et al. 2008).

12.3.2 Swap of Mutualist, Hybridization, or Novel Mutations?

Given the important role of the symbiont *Ishikawaella* in determining the pest-status of the insect in Asia, researchers questioned its role in the U.S. invasion, and

in particular, on the emergence of *M. cribraria* as a soybean pest in the first year. Could the emergence on soybeans be due to a "symbiont swap" in which the nonpest insect acquired the pest symbiont? *Ishikawaella*'s extracellular condition and capsule-transmission mode make this possibility feasible. In their study, Brown et al. (2013) sequenced the entire genomes of symbionts at 23 U.S. sample locations and compared these to the published symbiont genome from the pest *M. punctatissima* in Japan. They found only 47 fixed differences and only one of these (a riboflavin synthase subunit alpha, or *ribC* nonsynonymous change) appeared to have any potential to change symbiont nutrient-provisioning ability of the symbiont. This result strongly suggested that the U.S. *M. cribraria* had a symbiont that was nearly functionally identical to the pest-conferring symbiont in Japan. This was unexpected given the genetic and phenotypic resemblance of the U.S. insect to the nonpest in Asia (Hosokawa et al. 2005; Eger et al. 2010; Jenkins and Eaton 2011). The finding that these differences were fixed across the extant range in the U.S. including a specimen collected in the first week after discovery in 2009 ("ground zero" of the invasion), strongly suggested novel mutations in the U.S. were not responsible for the emergence of the pest phenotype. Instead, they suggest that the insect arrived already capable of being a soybean pest. The authors concluded that either a symbiont swap or hybridization or mixing of insects and symbionts may have occurred prior to arrival in the U.S. (Brown et al. 2013).

12.3.3 A Mutualist with Adaptive Potential: Microevolution

This same population genomics study (Brown et al. 2013) was also able to show possible evidence of adaptive potential in an obligate microbial mutualist in the form of polymorphic allele frequency change. The study showed that although there were no fixed mutations in U.S. *Ishikawaella* in the first 2 years (2009–2011), there were 164 low-frequency polymorphic sites at 1–28 % allele frequency, with up to 83 polymorphic sites per sample location. The majority of these mutations were new since "ground zero." By sequencing each sampling location at high depth of coverage ($\sim$500X), they were able to estimate allele frequency change comparing 2009 and 2011 across the genome. Overall, the polymorphic alleles showed a pattern consistent with purifying selection or population expansion (Tajima's D -2.6 to -4.4). While field data confirmed that the population was expanding dramatically, other genetic evidence supported purifying selection.

Frameshifts and nonsense mutations were significantly over-represented in the alleles that were decreasing in frequency since ground zero. Furthermore, the distribution of polymorphic alleles across the genome was nonrandom: change was clustered in a few COG functional categories (two-tailed likelihood-ratio test for goodness-of-fit to null of even distribution across COGs P-value $= 0.0049$, 16 df, after normalizing for the length of each COG). This supported the idea that a portion of the randomly occurring mutations had been purged. Collectively, these

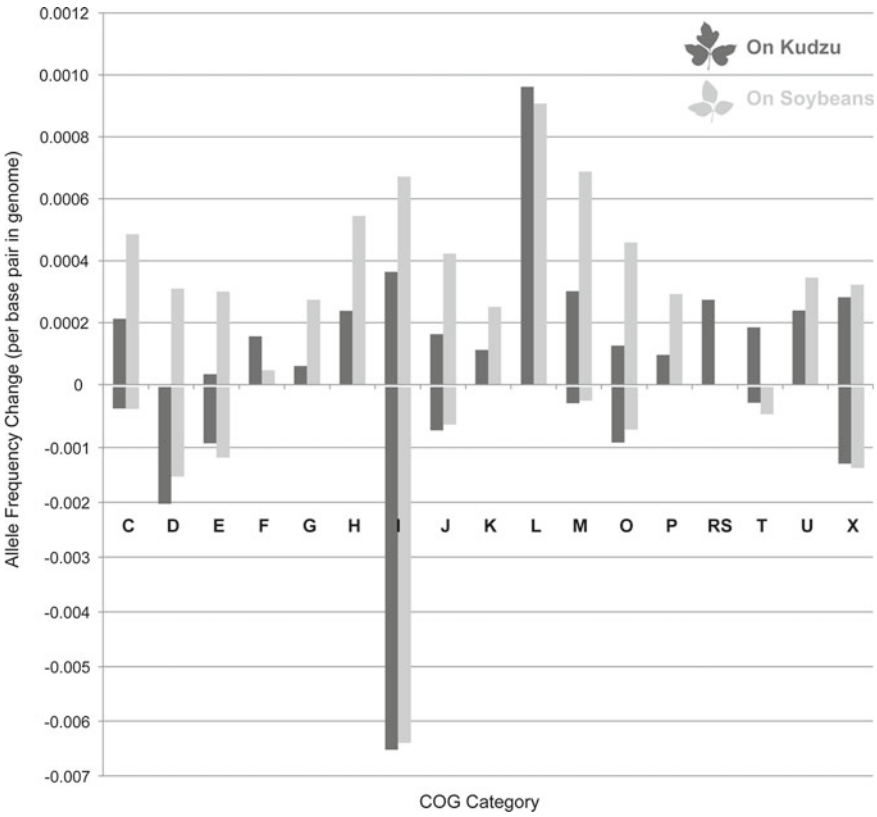


Fig. 12.3 Polymorphic allele frequency change in *Ishikawaella* compared to ground zero by functional group (COG category). Negative *axis* is condensed for easier visualization. *C* Energy production and conversion; *D* Cell cycle control, division, chromosome partitioning; *E* Amino acid transport and metabolism; *F* Nucleotide transport and metabolism; *G* Carbohydrate transport and metabolism; *H* Coenzyme transport and metabolism; *I* Lipid transport/metabolism; *J* Translation, ribosomal structure and biogenesis; *K* Transcription; *L* Replication, recombination and repair; *M* Cell wall/membrane/envelope biogenesis; *O* Post-translational modification, protein turnover, chaperones; *P* Inorganic ion transport and metabolism; *RS* general/unknown; *T* Signal transduction mechanisms; *U* Intracellular trafficking, secretion and vesicular transport. Adapted from Brown et al. (2013)

results support the idea that even during rapid expansion in which there might be relaxed selection, some purifying selection had indeed occurred.

Positive selection might appear as alleles with significant increase in frequency over time, but this could also occur by drift. However, positive selection could also be reflected by significant clustering of alleles in active sites of genes or in certain pathways. Both patterns were observed in *Ishikawaella*: statistical permutation tests showed significant over-representation of positive allele frequency change in several genes and pathways (Brown et al. 2013). Notably, this was statistically greater for samples from soybeans (see Fig. 12.3), wherein overall positive allele

frequency change on soybeans was greater than that on kudzu (two-tailed Wilcoxon signed rank test, $z_{crit} = -2.47$, $W = -105$, P-value $= 0.0135$) and especially for genes with probable symbiont role (amino acid and cofactor synthesis genes COGs E and H) with nonsynonymous change on soybeans 45.3 % versus on kudzu 7.2 %, one-tailed 2-sample unequal variance t-test P-value $= 0.016$) (Fig. 12.3) (Brown et al. 2013).

This study also helped to illuminate genes critical in the symbiont's role and potentially responsible for driving the success of its host in a new environment. For example, in the very short time since arrival of the insect (at that time of 2 years), mutational change was over-represented in genes and pathways that a priori were listed as candidates involved in "symbiont role," most notably, riboflavin synthesis. Riboflavin deaminase reductase (*ribD*) showed more mutations than expected in the catalytic Zn binding site and more positive change on soybeans (Brown et al. 2013). While it is unclear at the outset whether a symbiont polymorphism (e.g., an amino acid change in riboflavin synthase that is at less than 100 % frequency in a bacterial population within a single insect) could confer a difference in phenotype of the insect, results indirectly support that this may be the case.

Lastly, this study highlighted the potential to identify new genes or gene regions of interest. For example, a surprising finding was the significant increase in an allele in the recombination and repair gene Exonuclease V beta chain (*recB*) in a conserved region involved in helicase activity. This allele was positively correlated with new mutations elsewhere in the genome for samples on soybeans (Brown et al. 2013). If this relationship were causal, meaning that the *recB* allele causes an increase in genome-wide mutation, the implication could be that the allele would be rising in frequency due to stochastic effects, since most new mutations should be deleterious. Such as allele is unlikely to succeed unless driven on the wave of an expanding population, in a process called "surfing" (Travis et al. 2007; Excoffier and Ray 2008; François et al. 2010). See a map of locus-by-locus allele frequency change in Fig. 12.4.

12.4 Concluding Remarks and Prospects for Future Research

This chapter has highlighted the often overlooked potential role of microbial mutualists to influence the trajectory of evolution of their hosts. It outlined a natural experiment—a very recent invasive pest—in which a microbial mutualist with a limited genetic repertoire was shown to experience allele frequency change consistent with signatures of selection on the symbiont over a short timescale. What made this study possible was the approach: whole-genome deep sequencing of population data. This strategy has been successful in finding rare variants in HIV and somatic cancer cells (De Grassi et al. 2010) and has great promise in a

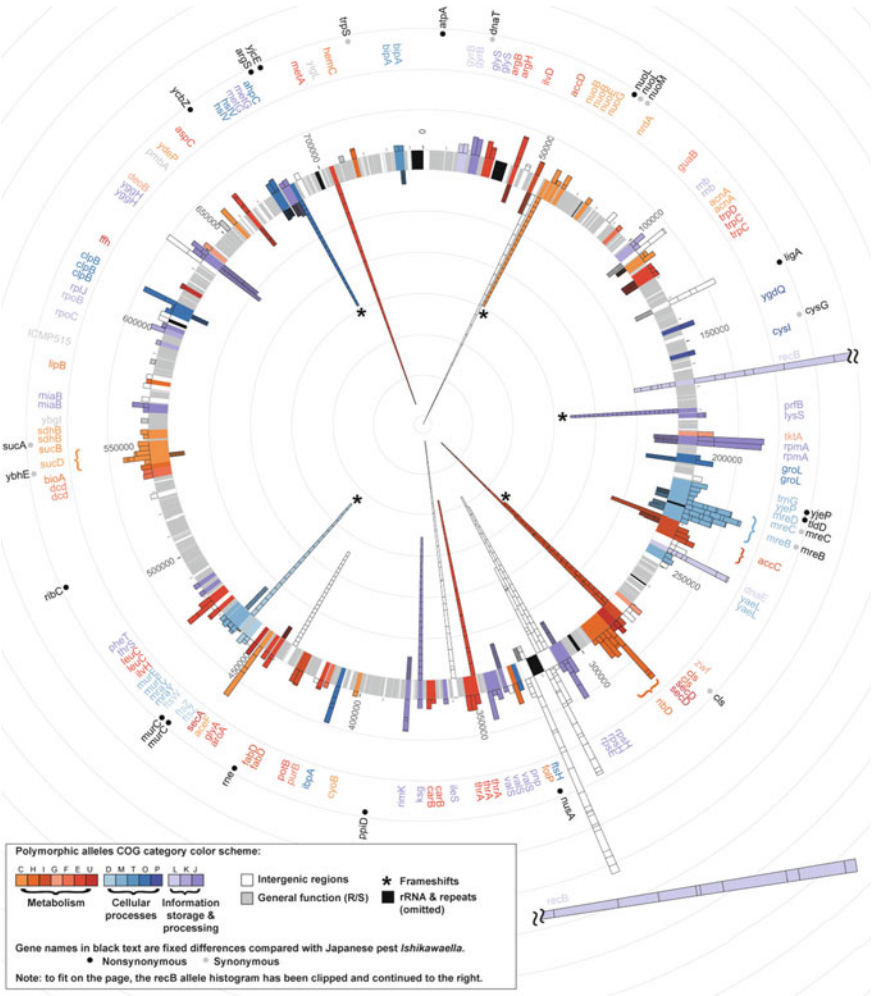


Fig. 12.4 Circular *Ishikawaella* genome showing additive allele frequency change since ground zero. Negative change is shown on *inward graph* and positive change is shown on *outward graph*. Adapted from Brown et al. (2013)

wide range of population studies (Barrett and Schluter 2008). The approach could be easily modified to achieve deep coverage and large population sizes for organisms with larger genomes by a targeting method such as FREQ-seq (Chubiz et al. 2012).

Acknowledgments I am grateful to J. P. McCutcheon and N. M. Gerardo for support in the population genomic study discussed in this chapter. I also thank L. Y. Huynh, C. M. Bolender, K. G. Nelson, T. M. Jenkins, D. R. Suiter, J. K. Greene, M. L. Allen, and J. T. Van Leuven for help. The latter was funded by a USDA AFRI grant (2011-67013-30090) to J. P. McCutcheon.

References

Abbot P, Moran NA (2002) Extremely low levels of genetic polymorphism in endosymbionts (*Buchnera*) of aphids (*Pemphigus*). Mol Ecol 11:2649–2660

Andersson DI (2008) Shrinking bacterial genomes. Microbe 3:124–130

Aslan CE, Zavaleta ES, Tershy B, Croll D (2013) Mutualism disruption threatens global plant biodiversity: a systematic review. PLoS ONE 8(6):e66993

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. Trends Ecol Evol 23:38–44

Baumann P (2005) Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol 59:155–189

Bennett AE (2013) Can plant–microbe–insect interactions enhance or inhibit the spread of invasive species? Funct Ecol 27(3):661–671

Bennett GM, Moran NA (2013) Small, smaller, smallest: the origins and evolution of ancient symbioses in a phloem-feeding insect. Genome Biol Evol 5(9):1675–1688

Brown A, Huynh LY, Bolender CM, Nelson KG, McCutcheon JP (2013) Population genomics of a symbiont in the early stages of a pest invasion. Mol Ecol (published online in advance of print, Jul 11). doi:10.1111/mec.12366

Buchner P (1965) Endosymbiosis of animals with plant microorganisms. Interscience Publishers, New York

Chubiz LM, Lee MC, Delaney NF, Marx CJ (2012) FREQ-Seq: a rapid, cost-effective, sequencing-based method to determine allele frequencies directly from mixed populations. PLoS ONE 7(10):e47959

Clay K, Holah J (1999) Fungal endophyte symbiosis and plant diversity in successional fields. Science 285:1742–1744

Dale C, Moran NA (2006) Molecular interactions between bacterial symbionts and their hosts. Cell 126:453–465

De Grassi A, Segala C, Iannelli F, Volorio S, Bertario L, Radice P, Bernard L, Ciccarelli FD (2010) Ultradeep sequencing of a human ultraconserved region reveals somatic and constitutional genomic instability. PLoS Biol 8:e1000275

Desprez-Loustau M-L, Robin C, Buée M, Courtecuisse R, Garbaye J, Suffert F, Sache I, Rizzo DM (2007) The fungal dimension of biological invasions. Trends Ecol Evol 22:472–480

Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. Trends Genet 27(4):157–163

Eger JE Jr, Ames LM, Suiter DR, Jenkins TM, Rider DA, Halbert SE (2010) Occurrence of the old world bug *Megacopta cribraria* (Fabricius) (Heteroptera: Plataspidae) in Georgia: a serious home invader and potential legume pest. Insecta Mundi 0121:1–11

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol Evol 23:347–351

Feldhaar H (2011) Bacterial symbionts as mediators of ecologically important traits of insect hosts. Ecol Entomol 36:533–543

Ferrari J, Vavre F (2011) Bacterial symbionts in insects or the story of communities affecting communities. Philos Trans R Soc B 366(1569):1389–1400

Forseth IN, Innis AF (2004) Kudzu (*Pueraria montana*): history, physiology, and ecology combine to make a major ecosystem threat. Crit Rev Plant Sci 23(5):401–413

François O, Currat M, Ray N, Han E, Excoffier L, Novembre J (2010) Principal component analysis under population genetic models of range expansion and admixture. Mol Biol Evol 27:1257–1268

Fukatsu T, Hosokawa T (2002) Capsule-transmitted gut symbiotic bacterium of the Japanese common plataspid stinkbug, *Megacopta punctatissima*. Appl Environ Microbiol 68:389–396

Funk DJ, Wernegreen JJ, Moran NA (2001) Intraspecific variation in symbiont genomes: bottlenecks and the aphid-Buchnera association. Genetics 157(2):477–489

Goodrich-Blair H, Hussa E (2013) It takes a village: ecological and fitness impacts of multipartite mutualism. Annu Rev Microbiol 67:161–178
Hansen AK, Moran NA (2013) The impact of microbial symbionts on host plant utilization by herbivorous insects. Mol Ecol (published online in advance of print, Aug 16). doi:10.1111/mec.12421
Hosokawa T, Kikuchi Y, Meng XY, Fukatsu T (2005) The making of symbiont capsule in the plataspid stinkbug *Megacopta punctatissima*. FEMS Microbiol Ecol 54:471–477
Hosokawa T, Kikuchi Y, Shimada M, Fukatsu T (2007) Obligate symbiont involved in pest status of host insect. Proc R Soc Lond Ser B 274:1979–1984
Hosokawa T, Kikuchi Y, Shimada M, Fukatsu T (2008) Symbiont acquisition alters behaviour of stinkbug nymphs. Biol Lett 4:45–48
Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson AC, von Dohlen CD, Fukatsu T, McCutcheon JP (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153(7):1567–1578
Janson EM, Stireman JO III, Singer MS, Abbot P (2008) Phytophagous insect–microbe mutualisms and adaptive evolutionary diversification. Evolution 62:997–1012
Jenkins TMJ, Eaton TD (2011) Population genetic baseline of the first plataspid stink bug symbiosis (Hemiptera: Heteroptera: Plataspidae) reported in North America. Insects 2:264–272
Herre EA, Knowlton N, Mueller UG, Rehner SA (1999) The evolution of mutualisms: exploring the paths between conflict and cooperation. Trends Ecol Evol 14(2):49–53
Kikuchi A, Kobayashi H (2010) Effect of injury by adult *Megacopta punctatissima* (Montandon) (Hemiptera: Plataspidae) on the growth of soybean during the vegetative stage of growth. Jpn J Appl Entomol Zool 54:37–43
Lichman E (2010) Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. Ecol Lett 13:1560–1572
MacDonald SJ, Thomas GH, Douglas AE (2011) Genetic and metabolic determinants of nutritional phenotype in an insect–bacterial symbiosis. Mol Ecol 20:2073–2084
McCutcheon JP, Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol 2:708–718
McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. Nature Rev Microbiol 10:13–26
Mira A, Moran NA (2002) Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. Microb Ecol 44(2):137–143
Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596
Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet 42:165–190
Nikoh N, Hosokawa T, Oshima K, Hattori M, Fukatsu T (2011) Reductive evolution of bacterial genome in insect gut environment. Genome Biol Evol 3:702–714
Nikoh N, McCutcheon JP, Kudo T, Miyagishima SY, Moran NA, Nakabachi A (2010) Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet 6(2):e1000827
Orphan VJ (2009) Methods for unveiling cryptic microbial partnerships in nature. Curr Opin Microbiol 12:231–237
Richardson DM, Allsopp N, D'Antonio CM, Milton SJ, Rejmanek M (2000) Plant invasions—the role of mutualisms. Biol Rev 75(1):65–93
Russell JA, Weldon S, Smith AH, Kim KL, Hu Y, Łukasik P, Doll S, Anastopoulos I, Novin M, Oliver KM (2013) Uncovering symbiont-driven genetic diversity across North American pea aphids. Mol Ecol 22(7):2045–2059
Sandström J, Pettersson J (1994) Amino acid composition of phloem sap and the relation to intraspecific variation in pea aphid (*Acyrthosiphon pisum*) performance. J Insect Physiol 40:947–955

Suiter DR, Eger JE Jr, Gardner WA, Kemerait RC, All JN, Roberts PM, Greene JK, Ames LM, Buntin GD, Jenkins TM, Douce GK (2010) Discovery and distribution of *Megacopta cribraria* (Hemiptera: Heteroptera: Plataspidae) in northeast Georgia. J Integr Pest Manage 1:1–4

Palmer TM, Doak DF, Stanton ML, Bronstein JL, Kiers ET, Young TP, Goheen JR, Pringle RM (2010) Synergy of multiple partners, including freeloaders, increases host fitness in a multispecies mutualism. Proc Nat Acad Sci 107(40):17234–17239

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecol Econ 52:273–288

Traveset A, Richardson DM (2011) Mutualisms: key drivers of invasions… key casualties of invasions. In: Rchardson DM (ed) Fifty years of invasion ecology: the legacy of Charles Elton. Wiley-Blackwell, Oxford, pp 143–160

Travis JM, Münkemüller T, Burton OJ, Best A, Dytham C, Johst K (2007) Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol Biol Evol 24:2334–2343

University of Georgia—Center for Invasive Species and Ecosystem Health (2013) Distribution map *Megacopta cribraria.* http://www.kudzubug.org/distribution_map.cfm. Accessed 20 Nov 2013

United States Department of Agriculture (2012) USDA Economic Research Service soybeans-oil-crops. http://www.ers.usda.gov/topics/crops/soybeans-oil-crops.aspx. Accessed 15 Oct 2012

Zhang Y, Hanula JL, Horn S (2012) The biology and preliminary host range of *Megacopta cribraria* (Heteroptera: Plataspidae) and its impact on kudzu growth. Environ Entomol 41:40–50

Chapter 13
Why Did Terrestrial Insect Diversity Not Increase During the Angiosperm Radiation? Mid-Mesozoic, Plant-Associated Insect Lineages Harbor Clues

Conrad Labandeira

Abstract Several studies provided evidence that family-level insect diversity remained flat throughout the initial mid-Cretaceous angiosperm radiation 125–90 million years ago. As this result has engendered considerable commentary, a reanalysis was done of a new dataset of 280 plant-associated insect families spanning the 174 million year interval of the Jurassic–Paleogene periods from 201 to 23 million years ago. Lineage geochronologic ranges were determined, and feeding attributes were characterized by: (i) dominant feeding guild (herbivore, pollinator, herbivore–pollinator, pollinator–mimic, xylophage); (ii) membership in one of eight functional feeding groups; and (iii) dominant plant host or host transition (cryptogam/fern only, cryptogam/fern → angiosperm, gymnosperm only, gymnosperm → angiosperm, angiosperm only). A time-series plot of insect lineages and their dominant plant–host affiliations resulted in four conclusions. First, insect lineages with dominant gymnosperm hosts reached a level of 95 families in the 35 million years preceding the initial angiosperm radiation. Second, earlier insect lineages with gymnosperm → angiosperm host transitions and newly originated insect lineages that developed dominant associations with emerging angiosperms rapidly diversified during the angiosperm radiation, later establishing a plateau of 110 families during a 20 million year interval after the initial angiosperm radiation. Third, these two diversity maxima were separated during the angiosperm radiation by a diversity minimum, the Aptian–Albian gap, indicating major turnover and time-lag effects associated with the extirpation and

C. Labandeira (✉)
Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, MRC-121, 37012, Washington, DC 20012, USA
e-mail: labandec@si.edu

C. Labandeira
Department of Entomology and BEES Program, University of Maryland, College Park, MD 20742, USA

C. Labandeira
College of Life Sciences, Capital Normal University, Beijing 100048, China

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_13,

acquisition of plant associations. Last, insect lineages most affected during this interval were herbivores and pollinators, exophagous feeders, and those hosting gymnosperms, angiosperms and gymnosperm → angiosperm transitions. These data largely explain the flat or even decreased level of insect diversity immediately before, during, and after the initial angiosperm radiation.

13.1 Introduction

In 1993, Labandeira and Sepkoski published a report in *Science* documenting the fossil diversity of insects from an assessment of their family-level lineage diversity through time. One of the results of this study was the clear presence of stasis in the rate of origination of insect diversity immediately before, during, and after the initial radiation of angiosperms in the mid-Cretaceous (Fig. 13.1). Some of the reaction to this discovery was negative, particularly from some paleoentomologists and paleobotanists. The discontent was attributable to the view that a slackening of insect diversity contravened a well-received view that the diversity of associated herbivores and pollinators should have significantly increased in concert with sharply increased diversification of angiosperm lineages. This "coevolutionary" view presumed that an expected, coordinated evolution would occur between angiosperms and their various insect associates in a multiplicative and opportunistic manner during an interval of resource expansion that would include food, shelter, mating sites, and other features essential for survival of insects and their host–plants. An unexpressed corollary to this view was that earlier gymnosperms were largely unavailable to insect lineages that potentially could interact with plants.

During the early 1990s to mid-2000s not much was known about the Mesozoic fossil record of insects and land plants. This ignorance affected understanding of gymnosperm relationships with insects prior to the major emergence of angiosperms during the Aptian to Turonian stages from 125 to 90 million years ago. There was limited evidence for the consumption of live plant tissues (Labandeira 2013), although a few studies documented several Eurasian, mid-Mesozoic insect lineages with pollen as gut contents (Krassilov et al. 2007). There was isolated documentation, overwhelmingly from paleobotanists, that evidence for insect herbivory was present on gymnospermous plants typical of mid-Mesozoic floras, such as puncture wounds on cheirolepidiaceous conifers (Watson 1977), borings in conifer woods (Zhou and Zhang 1989), and galls on bennettitalean foliage (Harris 1942), but these reports were too few to provide any convincing conclusion that preangiospermous floras had insects that used gymnosperms appreciably for food. Only recently have gymnosperm-dominated floras been systematically studied to document broad patterns of herbivory within specific habitat settings (Ding et al. 2014).

During this time, there were additional examinations of insect family-level diversity in the fossil record (Jarzembowski and Ross 1993, 1996; Alekseev et al. 2001).

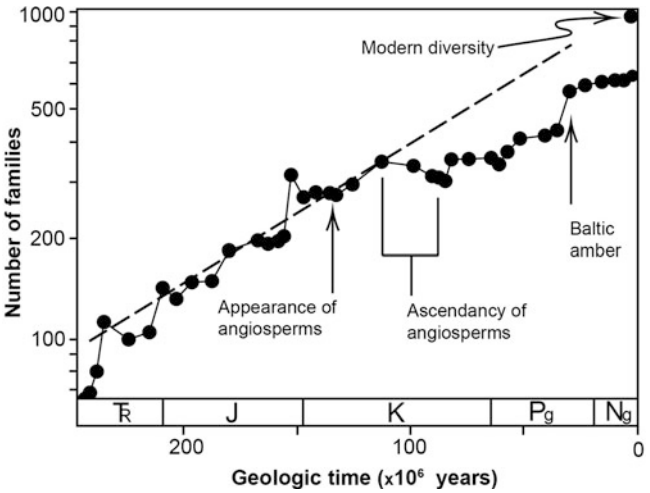


Fig. 13.1 A reproduction of Fig. 4 from Labandeira and Sepkoski (1993), showing insect family-level diversity increasing monotonically from the Middle Triassic to the Late Jurassic, stabilizing throughout the Cretaceous during the ascendancy of the angiosperms, and rising again during the Paleogene, indicated by the Baltic amber spike. The vertical axis is semilogarithmic and the dashed line is interpretive, indicating a divergence from exponential diversification during the Cretaceous. The only change to this figure has been relabeling of the abandoned "Tertiary Period" to reflect its modern division into an earlier Paleogene Period (Pg) and later Neogene Period (Ng). Reproduced with permission from American Association for the Advancement of Science

These reports confirmed that relative stasis existed for insect family-level diversity throughout the same interval of the angiosperm radiation, even though the data came from largely separately assembled datasets (Dmitriev and Zherikhin 1988; Rasnitsyn 1988; Labandeira 1994; Jarzembowski and Ross 1996). After publication of these studies and during the 2000s, a different approach was pursued—examination of the evidence for preangiospermous plant–insect interactions, focusing on herbivory, pollination, and eventually mimicry. Long-term projects with multiple colleagues were initiated to examine diverse preangiospermous, gymnosperm-dominated floras (Labandeira 2006; Ding et al. 2014). Current preliminary studies of Mesozoic herbivory use a similar methodology as those for the early Permian (Labandeira and Allen 2007; Schachat et al. 2014), the Cretaceous–Paleogene (K-Pg) boundary (Labandeira et al. 2002; Wilf et al. 2006), and the mid-Paleogene climate events (Wilf et al. 2005; Wappler et al. 2012; Currano et al. 2009). In addition, studies of mid-Mesozoic pollination have centered principally on the preangiospermous mid-Mesozoic (Ren et al. 2009; Labandeira 2010; Peñalver et al. 2012). Recently, there has been detection of mid-Mesozoic mimicry (Wang et al. 2012b). The purpose of these efforts in examining different aspects of plant–insect interactions and associations with gymnosperms in floras prior to the angiosperm radiation was to establish independent lines of evidence for

understanding why there was no increase in diversity commensurate with the angiosperm radiation.

In this contribution, a comprehensive summary of family-level, plant-associated lineages with dominantly cryptogam/fern, gymnosperm, and angiosperm host relationships is provided for an interval of time encompassing the Jurassic, Cretaceous, and Paleogene periods from 201 to 23 million years ago. The major focus of this report is to assess the diversity pattern and effects that the initial rise of angiosperms had on insect families that previously hosted cryptogams, ferns, and gymnosperms, and whether the subsequent shift of insect lineages toward angiosperm hosts is associated with a marked increase in their diversity. It is hoped that this exercise would spur other examinations of insect diversity patterns, such as assessments at the genus rank, through this formative time interval that resulted in much of the modern terrestrial world.

13.2 Methods and Definitions

13.2.1 Herbivory: Dominant Feeding Guilds, Functional Feeding Groups, and Plant Hosts

The first of these dietary habits of insects is the feeding guild (Table 13.1, Fig. 13.2). Five major feeding guilds are considered to encompass the variety of insect relationships with plants that are specified in the primary documentation (Table 13.1, Fig. 13.2). These dominant feeding guilds are *herbivory, pollination, herbivory–pollination, pollination–mimicry*, and *xylophagy*. The feeding guilds characterize family-level lineages of the eleven major orders of plant-associated insects during the Jurassic to Paleogene periods: Orthoptera, Phasmatodea, Thysanoptera, Hemiptera, Neuroptera, Coleoptera, Trichoptera, Lepidoptera, Mecoptera, Diptera, and Hymenoptera. These principal insect lineages constitute a broad variety of habitus types affiliated with orthopteroid, hemipteroid, and holometabolous developmental modes.

Herbivory is an antagonistic interaction defined as the consumption of live, photosynthetic plant tissues such as foliage, stems, and other organs. Although pollination is an interaction that may include the consumption of photosynthetic or non-photosynthetic tissues such as seeds, it is primarily characterized as the transfer of pollen from the pollinate to the ovulate reproductive organ of a conspecific plant host. When an insect taxon harbors two dominant plant-interactional strategies, such as an immature insect instar (nymph, larva) feeding on foliar tissues and an adult actively pollinating a different suite of host plants, such an interaction is herbivory–pollination. Similarly, a few pollinators possess the nonfeeding interaction of mimicry to deter or otherwise avoid insect predation but frequently are pollinators as well, in which case they are pollinator–mimics. Last, xylophagy is the consumption of wood, but for this study the feeding guild obligately includes the consumption of associated live tissue, such as subcortical

Table 13.1 Ranges and major habits of seed plant-associated family-level insect taxa during the Jurassic, Triassic and Cretaceous periods[a]

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Orthoptera (N = 16)					
Acrididae	J (Tith)–Recent	Herbivore	EF	Gymno → Angio	1
Elcanidae	P (Arti)–K (Apti)	Herbivore	EF	Gymno	2
Eumastacidae	J (Oxfo)–Recent	Herbivore	EF	Gymno → Angio	3
Haglidae	Tr (Anis)–Recent	Herb/Poll	EF	Gymno → Angio	4
Haglotettigoniidae	K (Albi)–K (Albi)	Herbivore	EF	Angio	5
Locustopseidae	Tr (Olen)–Pg (Chat)	Herbivore	EF	Gymno → Angio	6
Myrmecophilidae	K (Apti)–Recent	Herbivore	EF	Angio	7
Phasmomimidae	J (Toar)–Pg (Than)	Herbivore	EF	Gymno → Angio	8
Pneumoridae	K (Albi)–Recent	Herbivore	EF	Angio	9
Promastacidae	Pg (Than)–Pg (Than)	Herbivore	EF	Gymno	10
Prophalangopsidae	J (Hett)–Recent	Herb/Poll	EF	Gymno → Angio	11
Raphidiophoridae	K (Apti)–Recent	Herbivore	EF	Angio	12
Tetrigidae	K (Albi)–Recent	Herbivore	EF	Angio	13
Tettigoniidae	J (Sine)–Recent	Herbivore	EF	Gymno → Angio	14
Tuphellidae	Tr (Ladi)–J (Oxfo)	Herbivore	EF	Gymno	15
Vitimiidae	Tr (Ladi)–K (Albi)	Herbivore	EF	Gymno	16
Phasmatodea (N = 3)					
Phasmatidae	Pg (Pria)–Recent	Herbivore	EF	Angio	17
Phyllidae	Pg (Lute)–Recent	Herbivore	EF	Angio	18
Susmaniidae	J (Call)–Pg (Dani)	Herbivore	EF	Gymno → Angio	19
Thysanoptera (N = 11)					
Aeolopthripidae	K (Berr)–Recent	Herb/Poll	P&S	Gymno → Angio	20
Heterothripidae	K (Camp)–Recent	Herb/Poll	P&S	Angio	21
Karataothripidae	J (Oxfo)–J (Oxfo)	Herb/Poll	P&S	Gymno	22
Liassothripidae	J (Oxfo)–J (Oxfo)	Herb/Poll	P&S	Gymno	23
Lophioneuridae	P (Arti)–K (Albi)	Herb/Poll	P&S	Gymno	24
Melanthripidae	K (Albi)–Recent	Herb/Poll	P&S	Gymno → Angio	25
Merothripidae	K (Apti)–Recent	Herb/Poll	P&S	Angio	26
Moundthripidae	K (Apti)–K (Apti)	Herb/Poll	P&S	Gymno	27
Phlaeothripidae	K (Sant)–Recent	Herb/Poll	Gall	Angio	28
Stenurothripidae	K (Apti)–Recent	Herb/Poll	P&S	Angio	29
Thripidae	K (Apti)–Recent	Herb/Poll	P&S	Angio	30
Hemiptera (N = 103)					
Acanthosomatidae	Pg (Lute)–Recent	Herbivore	P&S	Angio	31
Adelgidae	K (Turo)–Recent	Herbivore	Gall	Gymno	32
Aleyrodidae	J (Oxfo)–Recent	Herbivore	P&S	Gymno → Angio	33
Alydidae	K (Haut)–Recent	Herbivore	P&S	Gymno → Angio	34
Anoecidae	Pg (Pria)–Recent	Herbivore	P&S	Angio	35
Anthocoridae	K (Berr)–Recent	Herbivore	SP	Gymno → Angio	36
Aphididae	J (Kimm)–Recent	Herbivore	P&S	Gymno → Angio	37
Aphrophoridae	Pg (Than)–Recent	Herbivore	P&S	Angio	38
Archescytinidae	C (Gzhe)–J (Toar)	Herbivore	P&S	Gymno	39
Archijassidae	Tr (Carn)–K (Apti)	Herbivore	P&S	Gymno	40
Berytidae	Pg (Pria)–Recent	Herbivore	P&S	Angio	41
Boreoscytidae	P (Kung)–K (Apti)	Herbivore	P&S	Gymno	42
Canadaphididae	K (Sant)–K (Camp)	Herbivore	P&S	Gymno	43

(continued)

Table 13.1 (continued)

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Carsidaridae	Pg (Pria)–Recent	Herbivore	P&S	Angio	44
Cercopidae	J (Call)–Recent	Herbivore	P&S	Gymno → Angio	45
Cercopionidae	Tr (Carn)–K (Apti)	Herbivore	P&S	Gymno	46
Cicadellidae	J (Kimm)–Recent	Herbivore	P&S	Gymno → Angio	47
Cicadidae	K (Ceno)–Recent	Herbivore	P&S	Angio	48
Cixiidae	Tr (Rhae)–Recent	Herbivore	P&S	Gymno → Angio	49
Clastopteridae	Pg (Pria)–Recent	Herbivore	P&S	Angio	50
Coccidae	Pg (Lute)–Recent	Herbivore	P&S	Angio	51
Coreidae	J (Call)–Recent	Herbivore	P&S	Gymno → Angio	52
Cretamyzidae	K (Camp)–K (Camp)	Herbivore	P&S	Gymno	53
Cuneocoridae	J (Plie)–J (Toar)	Herbivore	P&S	Gymno	54
Cydnidae	K (Berr)–Recent	Herbivore	P&S	Gymno → Angio	55
Dactylopiidae	Pg (Chat)–Recent	Herbivore	P&S	Angio	56
Delphacidae	Pg (Ypre)–Recent	Herbivore	P&S	Angio	57
Derbidae	Pg (Lute)–Recent	Herbivore	P&S	Angio	58
Diaspididae	K (Albi)–Recent	Herbivore	Gall	Angio	59
Dictyopharidae	K (Sant)–Recent	Herbivore	P&S	Angio	60
Drepanosiphidae	K (Turo)–Recent	Herbivore	P&S	Angio	61
Dunstaniidae	P (Road)–J (Sine)	Herbivore	P&S	Gymno	62
Dysmorphoptilidae	P (Arti)–J (Oxfo)	Herbivore	P&S	Gymno	63
Elektraphididae	K (Sant)–Ng (Piac)	Herbivore	P&S	Angio	64
Eoscarterellidae	P (Chan)–K (Berr)	Herbivore	P&S	Gymno	65
Eriococcidae	K (Turo)–Recent	Herbivore	Gall	Angio	66
Flatidae	Pg (Sela)–Recent	Herbivore	P&S	Angio	67
Fulgoridae	J (Toar)–Recent	Herbivore	P&S	Gymno → Angio	68
Fulgoridiidae	J (Hett)–L (Vala)	Herbivore	P&S	Gymno	69
Genaphididae	J (Call)–K (Tith)	Herbivore	P&S	Gymno	70
Granulidae	J (Call)–K (Apti)	Herbivore	P&S	Gymno	71
Hormaphididae	J (Kimm)–Recent	Herbivore	Gall	Gymno → Angio	72
Hylicellidae	P (Wuch)–J (Oxfo)	Herbivore	P&S	Gymno	73
Inkaidae	K (Sant)–K (Sant)	Herbivore	P&S	Angio	74
Ipsviciidae	P (Road)–J (Sine)	Herbivore	P&S	Gymno	75
Issidae	Pg (Lute)–Recent	Herbivore	P&S	Angio	76
Karabasiidae	J (Bath)–K (Haut)	Herbivore	P&S	Crypt/Fern	77
Karajassidae	J (Bath)–K (Apti)	Herbivore	P&S	Gymno	78
Kermesidae	Pg (Lute)–Recent	Herbivore	Gall	Angio	79
Kobdocoridae	K (Berr)–K (Apti)	Herbivore	P&S	Gymno	80
Lachnidae	Pg (Chat)–Recent	Herbivore	P&S	Gymno	81
Lalacidae	J (Call)–K (Apti)	Herbivore	P&S	Gymno	82
Largidae	Pg (Pria)–Recent	Herbivore	P&S	Angio	83
Latiscutellidae	K (Barr)–K (Apti)	Herbivore	P&S	Gymno	84
Liadopsyllidae	J (Toar)–K (Turo)	Herbivore	P&S	Gymno	85
Ligavenidae	K (Apti)–K (Apti)	Herbivore	P&S	Gymno	86
Lophophidae	J (Toar)–Recent	Herbivore	SP	Gymno → Angio	87
Lygaeidae	J (Bajo)–Recent	Herbivore	P&S	Gymno → Angio	88
Margarodidae	K (Haut)–Recent	Herbivore	Gall	Angio	89
Matsucoccidae	Pg (Lute)–Recent	Herbivore	P&S	Gymno	90

(continued)

Table 13.1 (continued)

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Membracidae	Pg (Lute)–Recent	Herbivore	P&S	Angio	91
Mesozoicoaphididae	K (Camp)–K (Camp)	Herbivore	P&S	Gymno	92
Mindaridae	K (Albi)–Recent	Herbivore	Gall	Gymno	93
Miridae	J (Oxfo)–Recent	Herbivore	P&S	Gymno → Angio	94
Neopsylloididae	J (Oxfo)–J (Oxfo)	Herbivore	P&S	Gymno	95
Nogodinidae	K (Berr)–Recent	Herbivore	P&S	Gymno → Angio	96
Ortheziidae	K (Apti)–Recent	Herbivore	P&S	Cryp/Fern → Angio	97
Oviparosiphidae	J (Toar)–K (Apti)	Herbivore	P&S	Gymno	98
Pachymeridiidae	Tr (Rhae)–K (Apti)	Herbivore	SP	Gymno	99
Palaeoaphididae	K (Berr)–K (Camp)	Herbivore	P&S	Gymno	100
Palaeontinidae	Tr (Ladi)–K (Apti)	Herbivore	P&S	Gymno	101
Pemphigidae	K (Sant)–Recent	Herbivore	Gall	Angio	102
Pentatomidae	K (Albi)–Recent	Herbivore	P&S	Angio	103
Pereboridae	P (Arti)–K (Barr)	Herbivore	P&S	Gymno	104
Phylloxeridae	K (Camp)–Recent	Herbivore	Gall	Angio	105
Piesmatidae	K (Albi)–Recent	Herbivore	P&S	Angio	106
Pityococcidae	Pg (Lute)–Recent	Herbivore	P&S	Angio	107
Pricecoridae	K (Barr)–Pg (Lang)	Herbivore	P&S	Angio	108
Procercopidae	J (Hett)–K (Barr)	Herbivore	P&S	Gymno	109
Progonocimicidae	P (Chan)–K (Apti)	Herbivore	P&S	Cryp/Fern	110
Protocoridae	J (Hett)–J (Hett)	Herbivore	P&S	Gymno	111
Protopsyllidiidae	P (Kung)–K (Apti)	Herbivore	P&S	Gymno	112
Pseudococcidae	Pg (Rupe)–Recent	Herbivore	Gall	Angio	113
Psyllidae	J (Oxfo)–Recent	Herbivore	Gall	Gymno → Angio	114
Pyrrhocoridae	Pg (Pria)–Recent	Herbivore	P&S	Angio	115
Rhopalidae	J (Call)–Recent	Herbivore	P&S	Gymno → Angio	116
Ricaniidae	Pg (Than)–Recent	Herbivore	P&S	Angio	117
Scutelleridae	Pg (Than)–Recent	Herbivore	P&S	Angio	118
Scytinopteridae	P (Kung)–K (Barr)	Herbivore	P&S	Gymno	119
Serpentivenidae	P (Word)–Pg (Sela)	Herbivore	P&S	Gymno → Angio	120
Shaposhnikoviidae	J (Bajo)–K (Sant)	Herbivore	P&S	Gymno	121
Sinojuraphidae	J (Call)–J (Call)	Herbivore	P&S	Gymno	122
Steingeliidae	K (Apti)–Recent	Herbivore	P&S	Angio	123
Stenoviciidae	P (Capi)–K (Berr)	Herbivore	P&S	Gymno	124
Tajmyraphididae	K (Apti)–K (Sant)	Herbivore	P&S	Gymno	125
Tettigarctidae	Tr (Carn)–Recent	Herbivore	P&S	Gymno → Angio	126
Tettigoniidae	Pg (Dani)–Recent	Herbivore	P&S	Angio	127
Thaumastocoridae	Pg (Ypre)–Recent	Herbivore	SP	Angio	128
Thelaxidae	K (Apti)–Recent	Herbivore	SP	Gymno	129
Tingidae	K (Berr)–Recent	Herbivore	Gall	Gymno → Angio	130
Velanthocoridae	J (Call)–K (Camp)	Pollinator	P&S	Gymno	131
Venicoridae	J (Call)–K (Barr)	Herbivore	P&S	Gymno	132
Weiwoboidae	Pg (Ypre)–Pg (Ypre)	Herbivore	P&S	Angio	133
Neuroptera (N = 4)					
Kalligrammatidae	J (Call)–K (Barr)	Poll/Mimic	SFF	Gymno	134
Nemopteridae	K (Apti)–Recent	Poll/Mimic	SFF	Angio	135
Staurosmylidae	J (Call)–J (Oxfo)	Poll/Mimic	SFF	Cryp/Fern	136
Panfiloviidae	J (Oxfo)–K (Apti)	Herbivore	SFF	Gymno	137

(continued)

Table 13.1 (continued)

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Coleoptera (N = 42)					
Aderidae	Pg (Lute)–Recent	Herbivore	EF	Angio	138
Anobiidae	K (Haut)–Recent	Xylophage	WB	Gymno → Angio	139
Attelabidae	K (Albi)–Recent	Herbivore	EF	Angio	140
Belidae	K (Barr)–Recent	Herb/Poll	Paly	Gymno	141
Boganiidae	J (Oxfo)–Recent	Pollinator	Paly	Gymno	142
Bostrichidae	J (Oxfo)–Recent	Xylophage	WB	Gymno → Angio	143
Brentidae	K (Barr)–Recent	Xylophage	WB	Angio	144
Bruchidae	K (Barr)–Recent	Herbivore	SP	Angio	145
Buprestidae	J (Call)–Recent	Xylophage	WB	Gymno → Angio	146
Byrrhidae	J (Oxfo)–Recent	Herbivore	EF	Cryp/Fern	147
Byturidae	Pg (Lute)–Recent	Pollinator	EF	Angio	148
Caridae	J (Oxfo)–Recent	Herb/Poll	EF	Gymno	149
Cerambycidae	K (Barr)–Recent	Xylophage	WB	Angio	150
Chrysomelidae	J (Call)–Recent	Herbivore	EF	Gymno → Angio	151
Curculionidae	J (Call)–Recent	Herb/Poll	EF	Gymno → Angio	152
Dascillidae	K (Apti)–Recent	Herbivore	EF	Angio	153
Erotylidae[g]	Pg (Ypre)–Recent	Herbivore	EF	Angio	154
Glaresidae	K (Barr)–Recent	Pollinator	EF	Angio	155
Ithyceridae	J (Call)–Recent	Herb/Poll	EF	Gymno → Angio	156
Lasiosynidae	J (Toar)–K (Apti)	Herb/Poll	EF	Gymno	157
Lymexylidae	K (Apti)–Recent	Xylophage	WB	Angio	158
Meloidae[h]	Pg (Lute)–Recent	Pollinator	SFF	Angio	159
Melyridae	K (Turo)–Recent	Pollinator	EF	Angio	160
Mordellidae	J (Call)–Recent	Pollinator	SFF	Gymno → Angio	161
Mycteridae	Pg (Ypre)–Recent	Pollinator	EF	Angio	162
Nemonychidae	J (Call)–Recent	Herb/Poll	Paly	Gymno	163
Nitidulidae	J (Call)–Recent	Herbivore	SFF	Gymno → Angio	164
Obrieniidae	Tr (Carn)–J (Oxfo)	Herbivore	EF	Gymno	165
Oedemeridae	Pg (Pria)–Recent	Herbivore	EF	Angio	166
Oxycorynidae	J (Oxfo)–Recent	Pollinator	SP	Gymno	167
Pandrexidae	J (Bajo)–K (Apti)	Herbivore	EF	Gymno	168
"Praemordellidae"[i]	J (Call)–J (Oxfo)	Pollinator	EF	Gymno	169
Protocucujidae	K (Apti)–Recent	Herbivore	EF	Angio	170
Protoscelidae	J (Call)–J (Oxfo)	Herbivore	EF	Gymno	171
Pythidae	Pg (Pria)–Recent	Xylophage	WB	Gymno	172
Ripiphoridae[j]	K (Albi)–Recent	Pollinator	SFF	Angio	173
Salpingidae	Pg (Ypre)–Recent	Herbivore	EF	Gymno	174
Scarabaeidae	J (Oxfo)–Recent	Pollinator	EF	Gymno → Angio	175
Scraptiidae	K (Berr)–Recent	Pollinator	EF	Gymno → Angio	176
Silvanidae	K (Albi)–Recent	Herbivore	SP	Angio	177
Ulyanidae	K (Apti)–K (Albi)	Herbivore	EF	Angio	178
Unnamed family	J (Call)–J (Call)	Pollinator	EF	Gymno	179
Trichoptera (N = 3)					
Dipseudopsidae	K (Turo)–Recent	Polllinator	SFF	Angio	180
Necrotauliidae[k]	Tr (Ladi)–K (Albi)	Pollinator	SFF	Gymno	181
Plectrotarsidae	J (Tith)–Recent	Pollinator	SFF	Gymno → Angio	182

(continued)

Table 13.1 (continued)

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Lepidoptera (N = 39)					
Adelidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	183
Agathiphagidae	K (Haut)[l]–Recent	Herb/Poll	SP	Gymno	184
Archaeolepidae	J (Sine)–J (Sine)	Herb/Poll	EF	Gymno	185
Ascololepidopterigidae	J (Call)–J (Call)	Herb/Poll	EF	Gymno	186
Bucculatricidae	K (Turo)–Recent	Herb/Poll	LM	Angio	187
Coleophoridae	K (Ceno)–Recent	Herb/Poll	LM	Angio	188
Copromorphidae	Pg (Pria)–Recent	Herb/Poll	EF	Angio	189
Cosmopterigidae	Pg (Chat)–Recent	Herb/Poll	LM	Angio	190
Cossidae	Pg (Chat)–Recent	Xylophage	WB	Angio	191
Elachistidae	Pg (Lute)–Recent	Herb/Poll	LM	Angio	192
Eolepidopterigidae	J (Call)–J (Oxfo)	Herb/Poll	EF	Gymno	193
Gelechiidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	194
Geometridae	Pg (Pria)–Recent	Herb/Poll	EF	Angio	195
Gracillariidae	K (Albi)–Recent	Herb/Poll	LM	Angio	196
Heliodinidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	197
Heliozelidae	Pg (Lute)–Recent	Herb/Poll	LM	Angio	198
Hesperiidae	Pg (Chat)–Recent	Herb/Poll	EF	Angio	199
Incurvariidae	Pg (Ypre)–Recent	Herb/Poll	LM	Angio	200
Lasiocampidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	201
Libytheidae	Pg (Pria)–Recent	Herb/Poll	EF	Angio	202
Lycaenidae	Pg (Chat)–Recent	Herb/Poll	EF	Angio	203
Lyonetiidae	Pg (Ypre)–Recent	Herb/Poll	LM	Angio	204
Micropterygidae	K (Apti)–Recent	Herb/Poll	Paly	Cryp/Fern → Angio	205
Mesokristenseniidae	J (Call)–J (Call)	Herb/Poll	EF	Gymno	206
Nepticulidae	K (Albi)–Recent	Herb/Poll	LM	Angio	207
Noctuidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	208
Nymphalidae	Pg (Ypr)–Recent	Herb/Poll	EF	Angio	209
Oecophoridae	Pg (Lute)–Recent	Xylophage	WB	Angio	210
Papilionidae	Pg (Ypre)–Recent	Herb/Poll	EF	Angio	211
Pieridae	Pg (Pria)–Recent	Herb/Poll	EF	Angio	212
Plutellidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	213
Psychidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	214
Pterophoridae	Pg (Chat)–Recent	Herb/Poll	EF	Angio	215
Pyralidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	216
Saturniidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	217
Thyrididae	Pg (Lute)–Recent	Xylophage	WB	Angio	218
Tortricidae	Pg (Lute)–Recent	Herb/Poll	EF	Angio	219
Undopterigidae[m]	J (Oxfo)–K (Apti)	Herb/Poll	LM[n]	Gymno	220
Zygaenidae	Pg (Chat)–Recent	Herb/Poll	EF	Angio	221
Mecoptera (N = 4)					
Aneuretopsychidae	J (Oxfo)–K (Apti)	Pollinator	SFF	Gymno	222
Cimbrophlebiidae	J (Toar)–Pg (Lute)	Poll/Mimic	EF	Gymno → Angio	223
Mesopsychidae	P (Wuch)–K (Barr)	Pollinator	SFF	Gymno	224
Pseudopolycentropodidae	Tr (Ladi)–K (Albi)	Pollinator	SFF	Gymno	225
Diptera (N = 27)					
Acroceridae	J (Oxfo)–Recent	Pollinator	SFF	Gymno → Angio	226
Agromyzidae	Pg (Dani)–Recent	Herbivore	LM	Angio	227
Anthomyiidae	Pg (Pria)–Recent	Pollinator	SFF	Angio	228

(continued)

Table 13.1 (continued)

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Anthomyzidae	Pg (Lute)–Recent	Herbivore	SFF	Angio	229
Apsilocephalidae	K (Albi)–Recent	Herbivore	SFF	Angio	230
Atelestidae	K (Berr)–Recent	Pollinator	Paly	Gymno → Angio	231
Athericidae	K (Berr)–Recent	Pollinator	SFF	Gymno → Angio	232
Bombyliidae	K (Berr)–Recent	Pollinator	Paly	Gymno → Angio	233
Cecidomyiidae	K (Haut)–Recent	Herbivore	Gall	Angio[o]	234
Chloropidae	Pg (Lute)–Recent	Herbivore	Gall	Angio	235
Cratomyiidae	K (Apti)–K (Apti)	Pollinator	Paly	Gymno	236
Hilarimorphidae	K (Haut)–Recent	Pollinator	SFF	Gymno → Angio	237
Lonchopteridae	K (Apti)–Recent	Pollinator	SFF	Gymno → Angio	238
Mydidae	K (Apti)–Recent	Pollinator	SFF	Gymno → Angio	239
Nemestrinidae	J (Call)–Recent	Pollinator	SFF	Gymno → Angio	240
Opomyzidae	Pg (Chat)–Recent	Herbivore	SFF	Angio	241
Platypezidae	K (Sant)–Recent	Herbivore	SFF	Angio	242
Protapioceridae[p]	J (Barr)–Recent	Pollinator	SFF	Angio	243
Scenopinidae	J (Oxfo)–Recent	Pollinator	SFF	Gymno → Angio	244
Stratiomyidae	K (Barr)–Recent	Pollinator	SFF	Angio	245
Syrphidae	K (Sant)–Recent	Pollinator	Paly	Angio	246
Tabanidae[q]	K (Berr)–Recent	Pollinator	SFF	Gymno → Angio	247
Therevidae	J (Oxfo)–Recent	Pollinator	SFF	Gymno → Angio	248
Tipulidae	Tr (Carn)–Recent	Pollinator	SFF	Gymno → Angio	249
Vermileonidae	J (Oxfo)–Recent	Pollinator	SFF	Gymno → Angio	250
Xylomyiidae	K (Albi)–Recent	Herbivore	SFF	Angio	251
Xylophagidae	K (Ceno)–Recent	Pollinator	WB	Angio	252
Hymenoptera (N = 28)					
Agaonidae	Pg (Rupe)–Recent	Pollinator	Gall	Angio	253
Anaxyelidae	J (Call)–Recent	Xylophage	WB	Gymno → Angio	254
Andrenidae	Pg (Lute)–Recent	Pollinator	SFF	Angio	255
Apidae	K (Maas)–Recent	Pollinator	SFF	Angio	256
Argidae	Pg (Than)–Recent	Herbivore	EF	Angio	257
Blasticotomidae	Pg (Lute)–Recent	Herbivore	WB	Cryp/Fern	258
Cephidae	J (Bajo)–Recent	Herbivore	EF	Gymno → Angio	259
Chalcididae	Pg (Ypre)–Recent	Herbivore	Gall	Angio	260
Cimbicidae	Pg (Than)–Recent	Herbivore	EF	Angio	261
Cynipidae	K (Camp)–Recent	Pollinator	Gall	Angio	262
Diprionidae	Pg (Rupe)–Recent	Herbivore	EF	Gymno	263
Electrotomidae	Pg (Lute)–Pg (Lute)	Herbivore	EF	Angio	264
"Gigasiricidae"[i]	K (Plie)–K (Apti)	Xylophage	WB	Gymno	265
Halictidae	Pg (Ypre)–Recent	Pollinator	SFF	Angio	266
Masaridae	K (Albi)–Recent	Pollinator	Paly	Angio	267
Megachilidae	Pg (Than)–Recent	Pollinator	SFF	Angio	268
Melittidae	Pg (Ypre)–Recent	Pollinator	SFF	Angio	269
Melittosphecidae	K (Albi)–K (Albi)	Pollinator	SFF	Angio	270
Pamphiliidae	J (Oxfo)–Recent	Herbivore	EF	Gymno → Angio	271
"Praesiricidae"[i]	J (Call)–K (Apti)	Herbivore	EF	Gymno	272
Scoliidae	J (Kimm)–Recent	Pollinator	SFF	Gymno → Angio	273
Sepulcidae	J (Plie)–K (Ceno)	Herbivore	EF	Gymno	274
Siricidae	J (Plie)–Recent	Xylophage	WB	Gymno → Angio	275
Tenthredinidae	J (Kimm)–Recent	Herbivore	LM	Cryp/Fern → Angio	276

(continued)

Table 13.1 (continued)

Order, family[b]	Geochronologic range[c]	Dominant feeding guild[d]	Major FFG[e]	Dominant hosts and host transitions[f]	Fig. 13.2 entry
Xiphydriidae	K (Albi)[12]–Recent	Xylophage	WB	Angio	277
Xyelidae	Tr (Ladi)–Recent	Pollinator	Paly	Gymno	278
"Xyelotomidae"[i]	J (Oxfo)–K (Barr)	Herbivore	EF	Cryp/Fern	279
"Xyelydidae"[i]	J (Toar)–K (Barr)	Herbivore	EF	Gymno	280

Notes

[a] Major data sources for this table are: Miller (1956), Lewis (1973), McAlpine et al. (1981, 1987, 1989), Gauld and Bolton (1988), Dmitriev and Zherikhin (1988), Naumann et al. (1991), Dolling (1991), Carpenter (1992), Ross and Jarzembowski (1993), Goulet and Huber (1993), Labandeira (1994), Evenhuis (1994); Schuh and Slater (1995), Rasnitsyn and Quicke (2002), Grimaldi and Engel, 2005), Yeates and Wiegmann (2005), Ren (2010), Sohn et al. (2012), Marshall (2012), Lawrence and Ślipiński (2013) and the Paleobiology Database (PBDB, 2014). Entries for ecological data categories indicate major attributes for the indicated taxon

[b] Depending on the authority, several of these family-level taxa (N = 280) have been reassigned to subfamily rank. Other family-level designations are currently in flux and require revision

[c] Abbreviations: The geologic periods, from oldest youngest, are: *C*, Carboniferous; *P*, Permian; *Tr*, Triassic; *J*, Jurassic; *K*, Cretaceous; *Pg*, Paleogene; and *Ng*, Neogene. The geologic stages, from oldest to youngest are the following. For the Carboniferous: *Gzhe*, Gzhelian. For the Permian: *Arti*, Artinskian; *Kung*, Kungurian; *Road*, Roadian; *Word*, Wordian; *Capi*, Capitanian; *Wuch*, Wuchiapingian; *Chan*, Changhsingian. For the Triassic: *Olen*, Olenekian; *Anis*, Anisian; *Ladi*, Ladinian; *Carn*, Carnian; *Rhae*, Rhaetian. For the Jurassic: *Hett*, Hettangian; *Sine*, Sinemurian; *Plie*, Pliensbachian; *Toar*, Toarcian; *Bajo*, Bajocian; *Bath*, Bathonian; *Call*, Callovian; *Oxfo*, Oxfordian; *Kimm*, Kimmeridgian; *Tith*, Tithonian. For the Cretaceous: *Berr*, Berriasian; *Vala*, Valanginian; *Haut*, Hauterivian; *Barr*, Barremian; *Apti*, Aptian; *Albi*, Albian; *Ceno*, Cenomanian; *Turo*, Turonian; *Sant*, Santonian; *Camp*, Campanian; *Maas*, Maastrichtian. For the Paleogene: *Dani*, Danian; *Sela*, Selandrian; *Than*, Thanetian; *Ypre*, Ypresian; *Lute*, Lutetian; *Pria*, Priabonian; *Rupe*, Rupelian; *Chat*, Chattian. For the Neogene: *Piac*, Piacenzian

[d] Abbreviations for the dominant feeding guilds of herbivory, pollination, xylophagy and mimicry are: *Herbivore*, *Pollinator*, *Herb/Poll*, *Poll/Mimic*, and *Xylophage* indicates the most dominant or otherwise most important feeding guild within the group under consideration

[e] Abbreviations for major functional feeding groups (*FFGs*) are: *EF*, external feeding; *Gall*, Galling; *LM*, leaf mining; *Paly*, pollen or spore consumption; *P&S*, piercing and sucking; *SFF*, surface fluid feeding; *SP*, seed predation; *WB*, wood boring

[f] Abbreviations for dominant hosts and transitions: *Cryp/Fern*, Cryptogams and/or ferns; *Gymno*, Gymnosperms; *Angio*, Angiosperms; *Cryp/Fern* → *Angio*, cryptogam or fern to angiosperm transition; *Gymno* → *Angio*, gymnosperm-to-angiosperm transition

[g] These pleasing fungus beetles include the Languriidae, the taxon of interest

[h] The relevant taxon is the Nemognathinae, with specialized, anthophilous mouthpart features

[i] These are stem groups possessing features of several derived taxa, and are considered here as distinctive lineages

[j] The relevant taxon is *Macrosiagon* and related genera, which possess specialized, anthophilous mouthpart features

[k] Probably more appropriately assigned to the Aphiesmenoptera, as a stem lineage to the Trichoptera + Lepidoptera

[l] Probable occurrence

[m] This family is not recognized by Sohn et al. (2012)

[n] Inferred larval feeding pattern

[o] Evidence suggests that the basal clades of Cecidomyiidae were not gall-forming, and that the gall-forming life habit evolved after angiosperms were established

[p] The Protapioceridae may be confamilial with the modern Apioceridae; alternatively, this lineage also may be ancestral to the Mydidae and Apioceridae. It is considered a distinct lineage herein

[q] The plant-associated subfamily, Pangioninae, is the relevant subgroup of Tabanidae that is considered herein

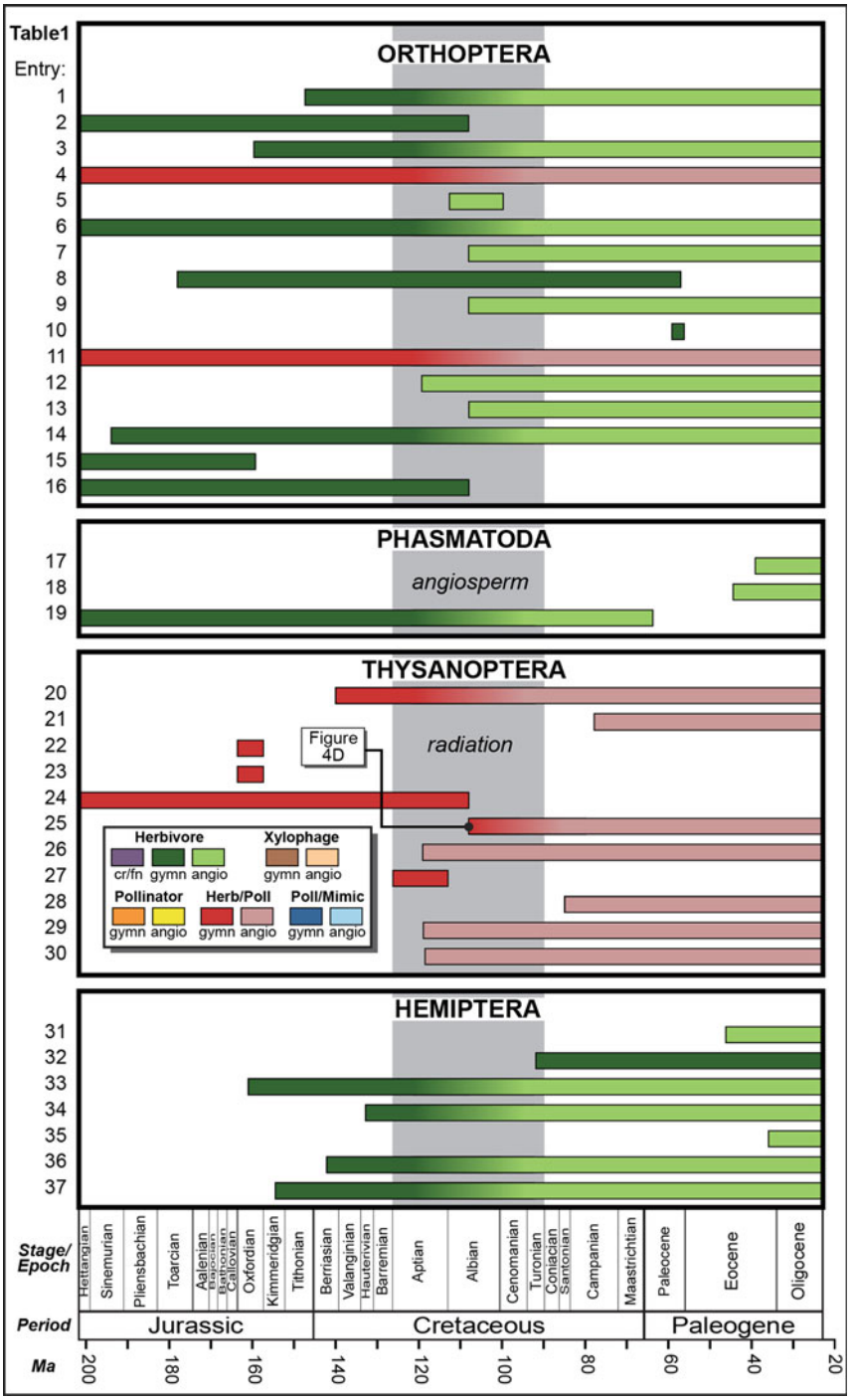


Fig. 13.2 (continued)

◀ **Fig. 13.2** Distribution of the eleven major, plant-associated insect lineages and their host–plant assignments during the Jurassic to Paleogene interval. The 35 million year-long angiosperm radiation encompasses the Aptian through Turonian stages of the mid Cretaceous as a vertical gray column at center. Major plant–host associations of herbivory, pollination, xylophagy, herbivory–pollination, and herbivory–mimicry, and their dominance in cryptogam/fern (*cr/fn*), cryptogam/fern → angiosperm, gymnosperm (*gymno*), gymnosperm → angiosperm and angiosperm (*angio*) hosts are indicated in the inset. Darker hues indicate gymnosperm hosts; lighter hues indicate angiosperm hosts. Data are from Table 13.1

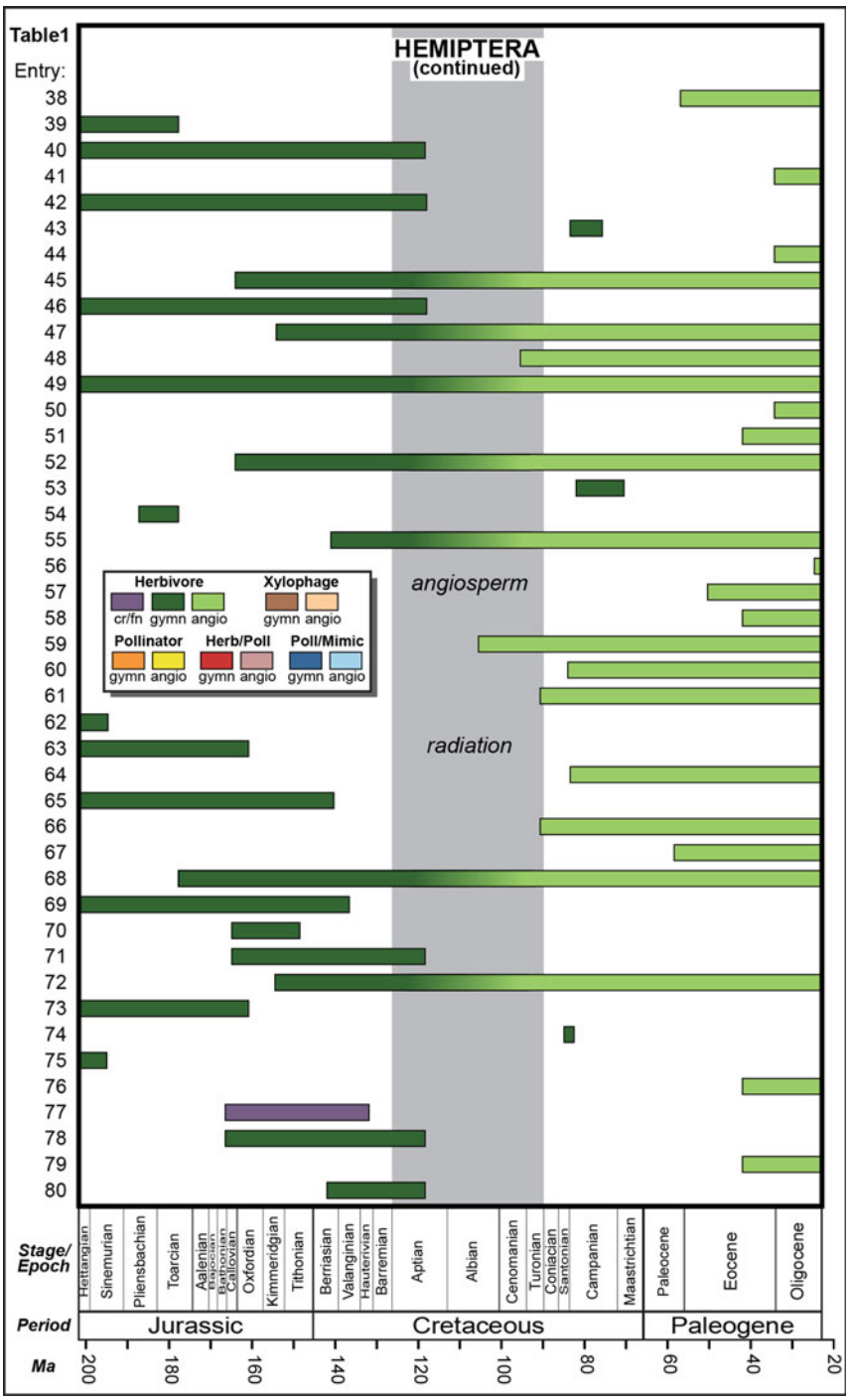


Fig. 13.2 (continued)

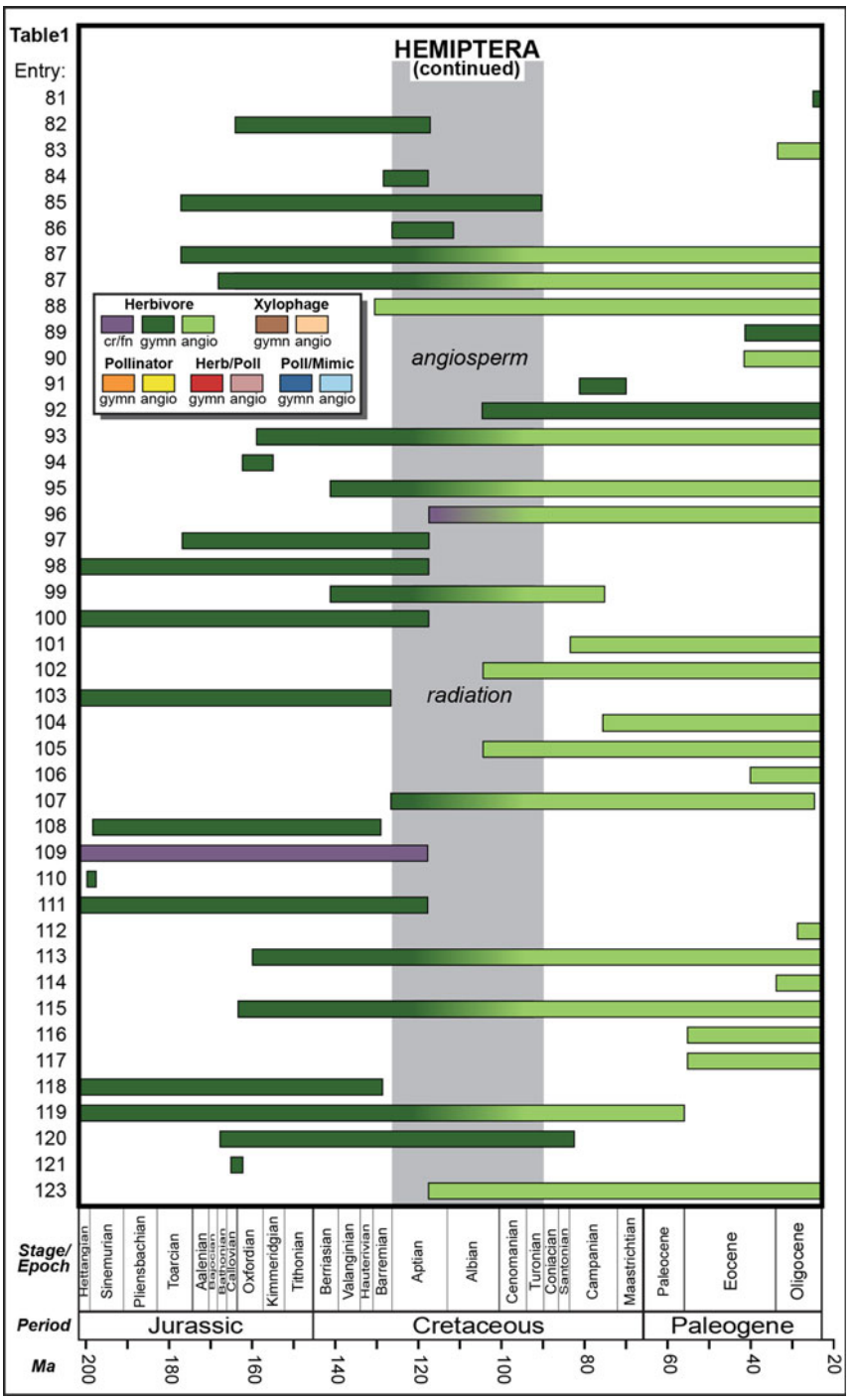


Fig. 13.2 (continued)

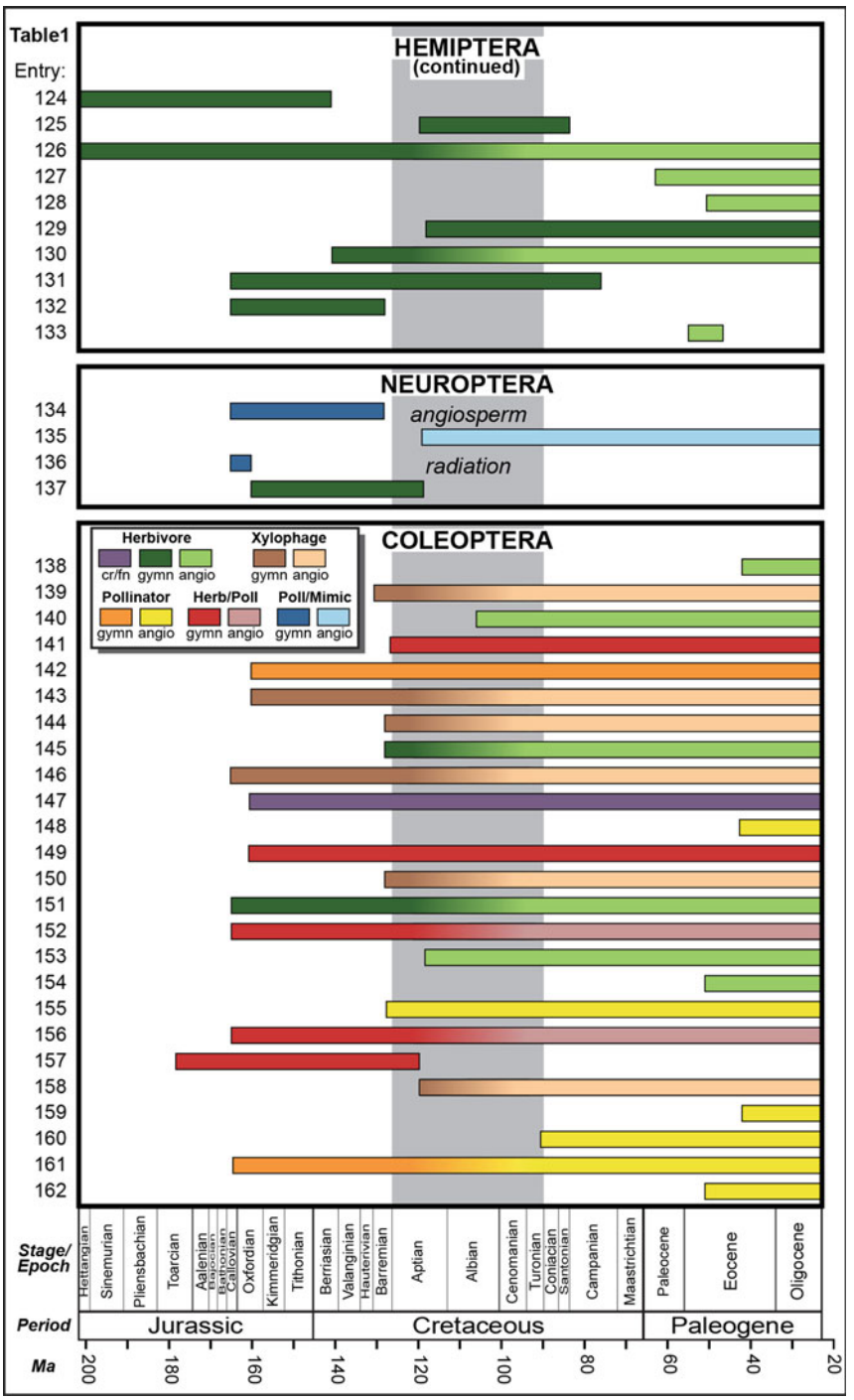


Fig. 13.2 (continued)

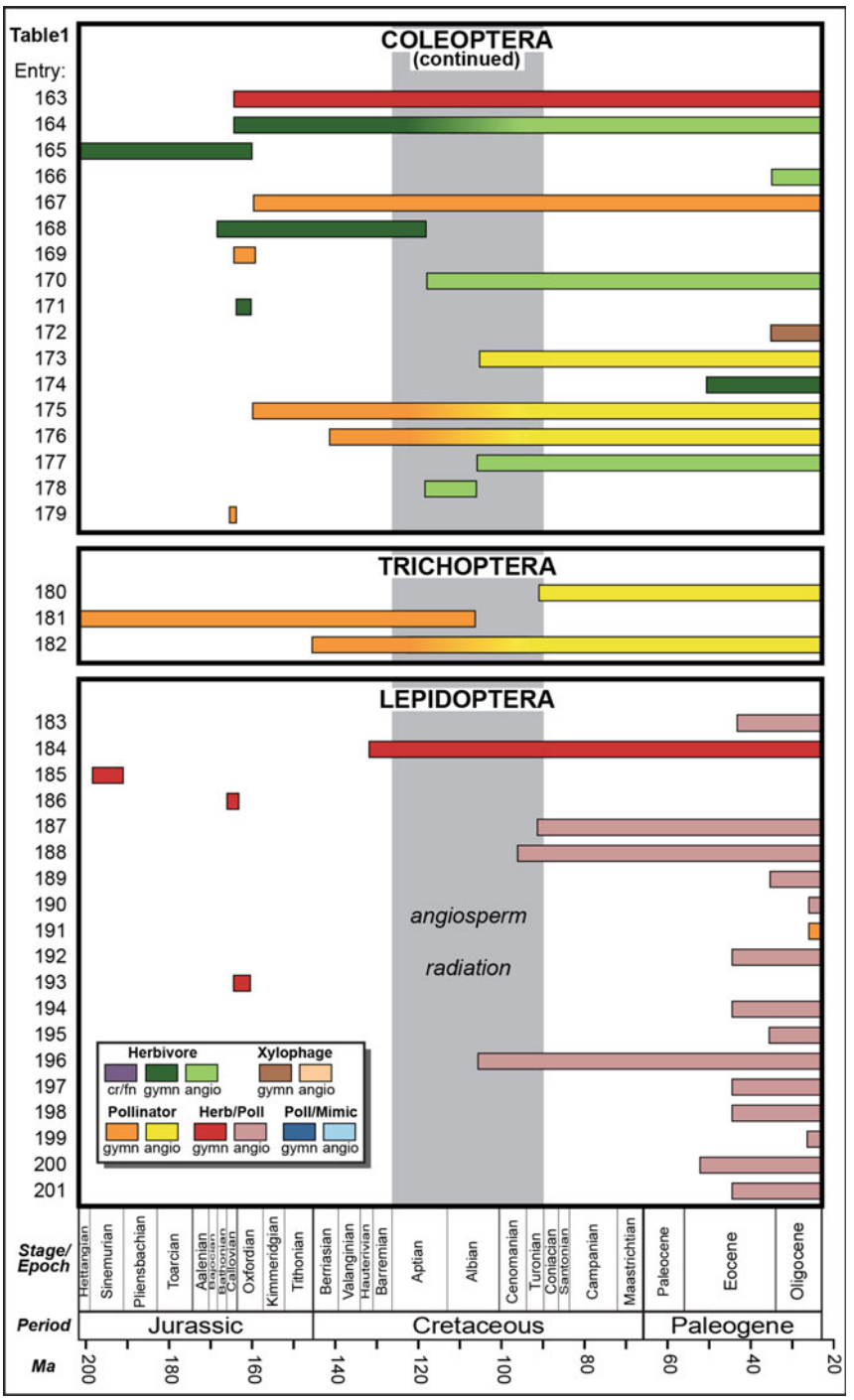


Fig. 13.2 (continued)

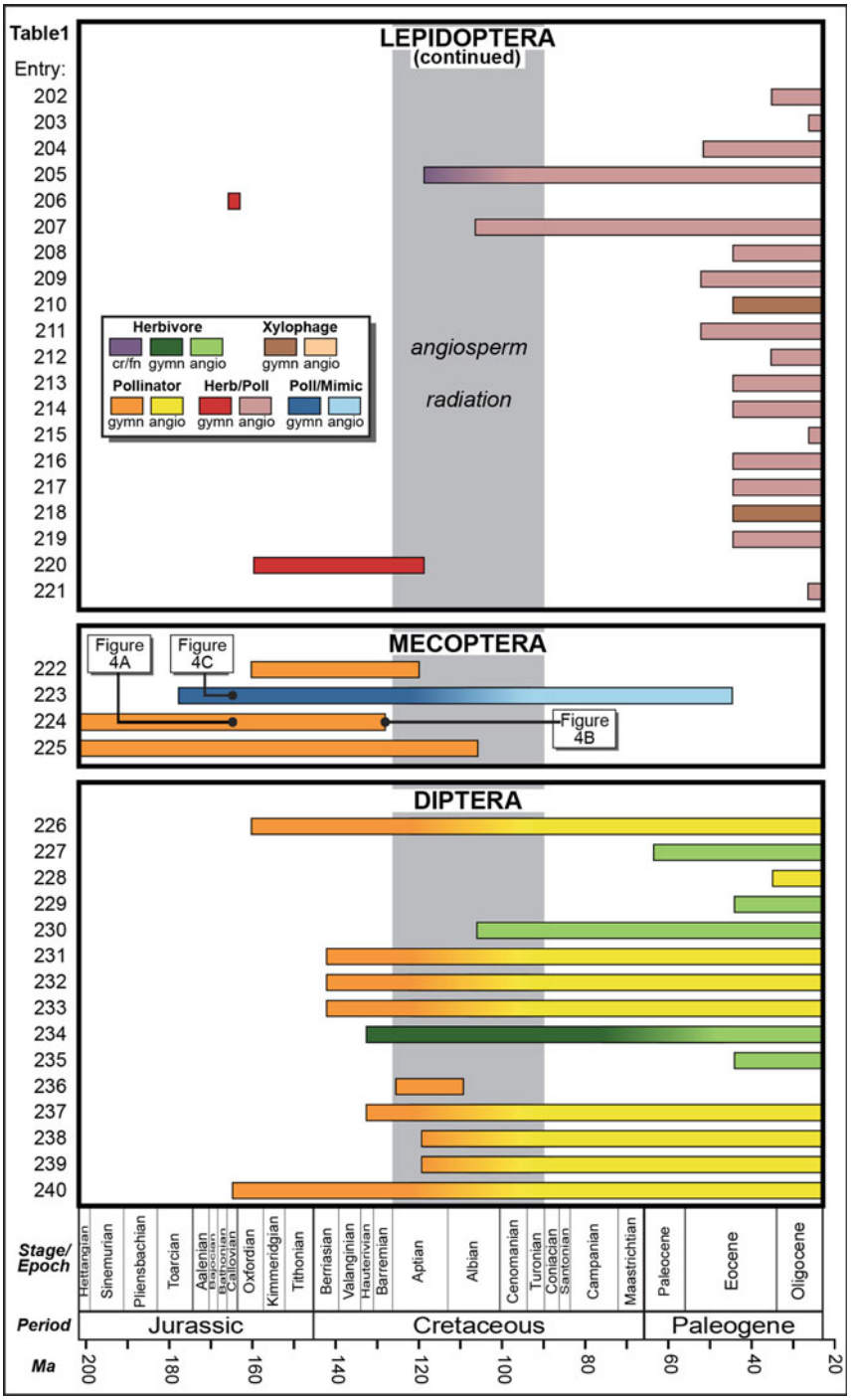


Fig. 13.2 (continued)

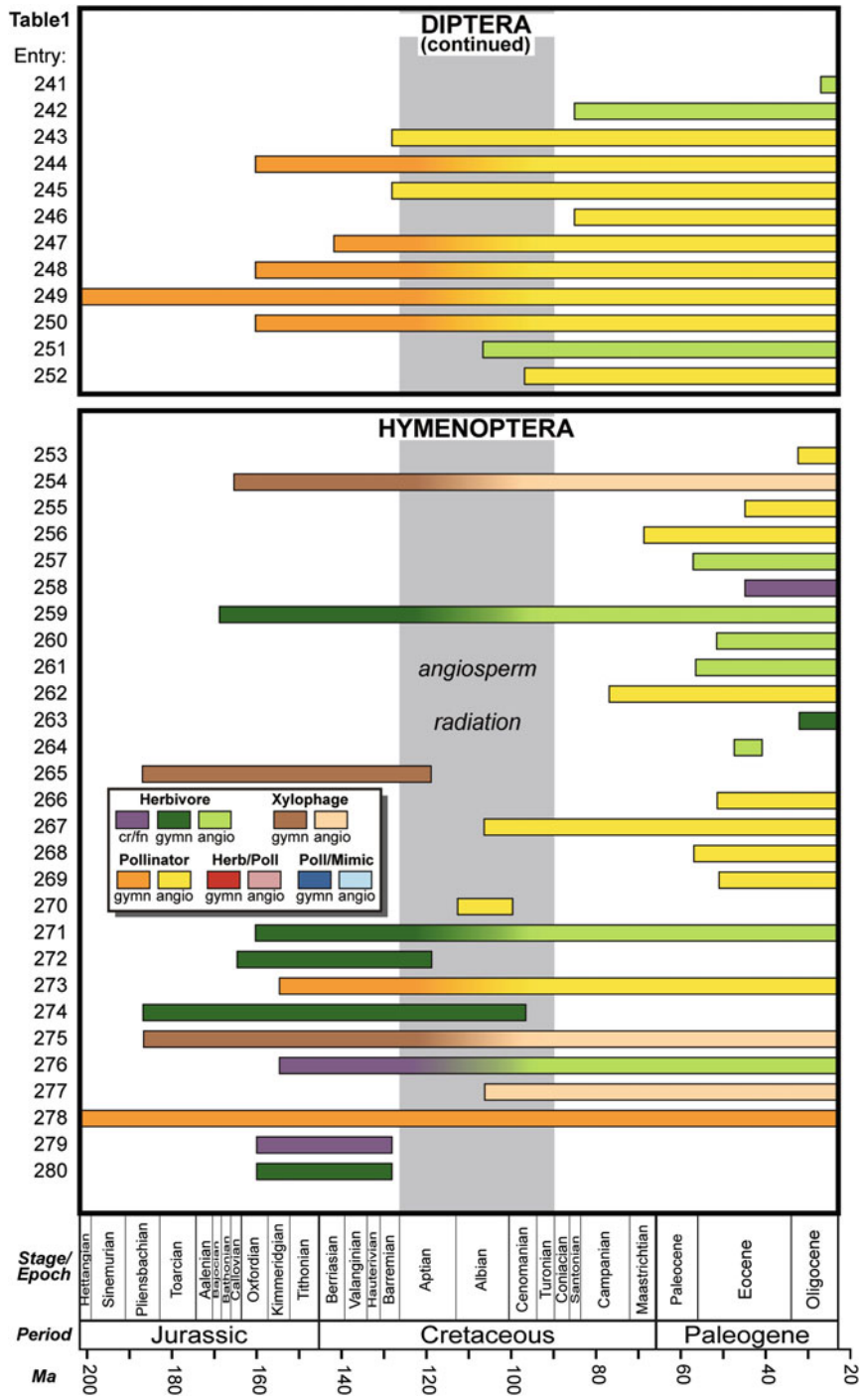


Fig. 13.2 (continued)

cambia, pith parenchyma, or other meristematic tissues that include live, actively dividing cells.

The second dietary habit is the functional feeding group. The data are divided into eight functional feeding groups for a more discrete, ecologically different characterization (Tables 13.1 and 13.2). The functional feeding groups are the modes of access to food that are effected principally through the action of mouthparts. *External feeding* is the consumption of foliage such as skeletonization and margin feeding in which the insect is outside of the tissue being consumed. *Piercing and sucking* consists of puncturing host tissues by specialized, stylate mouthparts and the subsequent sucking of fluid food. *Surface fluid feeding* is where surface fluids, such as pollination drops, floral or extrafloral nectar, or other plant exudates, are imbibed without inflicting a wound. *Palynivory*, or consumption of pollen, can be achieved by a variety of insect mouthpart types, in which ingestion may represent punctured pollen grains or entire to highly fragmented pollen clusters. Most pollinating insects are surface fluid feeders or palynivores.

The four previous functional feeding groups are ectophagous, occurring with the insect to the outside of the tissue consumed; by contrast, the following four interactions are endophytic, whereby the insect, typically an immature such as a larva or nymph, is lodged within plant tissues. *Galling* is a complex interaction whereby an insect immature inhabits a chamber surrounded by tumor-like plant tissues of newly created, inner, nutritive tissue for larval sustenance, outer hardened tissues for protection, and vascular tissue for food and water supply. Gall interactions are essentially parasitic and the galler arthropod hormonally controls the plant–host tissue and organ development adjacent the gall. *Leaf mining* is another endophytic interaction wherein an egg hatches into a larva that begins to consume foliage tissue, leaving a distinct frass trail and a leaf-mine with features such as successive width enlargements and a terminal chamber often used for pupation. *Seed predation* represents a variety of herbivore feeding types that have the common effect of consuming the embryonic and sustaining tissues of an ovule or seed. *Wood boring* consists of consumption of live meristematic tissues and parenchyma, often associated with tunneling through wood, the fabrication of borings, galleries and pupal chambers, and the consumption of associated fungi.

The third major characterization of dietary attributes of insect lineages in Table 13.1 is the dominant plant hosts and host transitions, linked to Fig. 13.2. We present five principle hosts and host transitions during the 174 million year interval from the Triassic–Jurassic to the Paleogene–Neogene boundaries. First, some insect lineages targeted cryptogam or fern hosts (*cryp/fern*) only. Second, other cryp/fern insect lineages have transitioned from cryp/fern to angiosperm hosts (*cryp/fern* → *angio*). Third, many insect lineages have always had dominantly gymnosperm hosts in their history. Fourth, some insect lineages on gymnosperms have switched their dominant hosts to angiosperms (*gymno* → *angio*). Fifth, more recent insect lineages have always had angiosperms (*angio*) as their dominant hosts.

Cryptogams included the familiar groups of liverworts, mosses, and lycopods; ferns consist of horsetails and marattialean and filicalean ferns. By contrast, gymnosperms include a diverse spectrum of extinct lineages (Taylor et al. 2009),

Table 13.2 Ecological attributes of the eleven insect lineages associated with the Mid-Mesozoic gymnosperm-to-angiosperm transition

Lineage examined	Dominant feeding guild					Major functional feeding group[a]								Dominant hosts and transitions[b]				
	Herbivore	Herb/ Poll	Pollinator	Poll/ Mimic	Xylophage	EF	P&S	Gall	SP	SFF	WB	Paly	LM	A Cryp/ Fern	B Cryp/ Fern → Angio	C Gymno only	D Gymno → Angio	E Angio only
Orthoptera	14	2	–	–	–	16	–	–	–	–	–	–	–	–	–	4	7	5
Phasmatodea	3	–	–	–	–	3	–	–	–	–	–	–	–	–	–	–	1	2
Thysanoptera	–	11	–	–	–	–	10	1	–	–	–	–	–	–	–	4	2	5
Hemiptera	103	–	–	–		–	87	12	4	–	–	–	–	2	1	43	22	35
Neuroptera	1	–	–	3	–	–	–	–	–	4	–	–	–	1	–	1	–	1
Coleoptera	16	6	13	–	7	25	–	–	3	4	7	3	–	1	–	13	14	14
Trichoptera	–	–	3	–	–	–	–	–	–	3	–	–	–	–	–	1	1	1
Lepidoptera	–	36	–	–	3	24	–	–	1	–	3	1	10	–	1	6	–	32
Mecoptera	–	–	3	1	–	1	–	–	–	3	–	–	–	–	–	3	1	–
Diptera	8	–	19	–	–	–	–	2	–	19	1	4	1	–	–	1	15	11
Hymenoptera	13	–	11	–	4	10	–	3	–	7	5	2	1	2	1	6	5	14
Totals	158	55	49	4	14	79	97	18	8	40	16	10	12	4	3	82	68	120

[a] Abbreviations for major functional feeding groups (*FFGs*) are: *EF*, external feeding; *Gall*, galling; *LM*, leaf mining; *Paly*, pollen or spore consumption; *P&S*, piercing and sucking; *SFF*, surface fluid feeding; *SP*, seed predation; *WB*, wood boring

[b] Abbreviations for dominant hosts and transitions: *Cryp/Fern*, cryptogams and/or ferns; *Gymno*, gymnosperms; *Angio*, angiosperms; *Cryp/Fern → Angio*, cryptogam or fern to angiosperm transition; *Gymno → Angio*, gymnosperm-to-angiosperm transition

such conifers, caytonialean and corystospermalean seed ferns, diverse ginkgophytes, bennettitaleans, and pentoxylaleans. Most gymnosperm sublineages became extinct during the angiosperm radiation, although several lineages now are known to have survived into the Gondwanan Paleogene, such as cheirolepidiaceous conifers (Barreda et al. 2012), corystosperm seed ferns (McLoughlin et al. 2008), Mesozoic-style ginkgoaleans (Hill and Carpenter 1999), and bennettitaleans (McLoughlin et al. 2011). By the close of the angiosperm radiation, all major groups of angiosperm lineages were established, including basal "paleoherb" lineages, monocots, Chloranthaceae, eumagnoliids, and core eudicots (Friis et al. 2010), and achieved ecological prominence in local habitats (Crane 1987).

13.2.2 Data Collection

Several initial conventions were used to provide a chronology of the summarized data (Table 13.1, Fig. 13.2). The 174 million year interval from the Triassic–Jurassic boundary at 201 Ma to the Paleogene–Neogene boundary at 23 Ma was used to document time durations of all identified plant-associated lineages. This time interval, consisting of the Jurassic, Cretaceous, and Paleogene periods, is divided into ca. 75 million years before the beginning of the angiosperm radiation at 125 Ma, and ca. 75 million years after its end at 90 Ma, providing a sufficiently long interval to record lineage turnover, long-term host–plant associations and major host transitions. These host associations occurred during the 35 million year-long angiosperm radiation from 125 Ma (ca. Barremian–Aptian boundary) to 90 Ma (ca. Turonian–Coniacian boundary). Insect lineage occurrence data were plotted at the midpoint for each geologic stage in which the insect lineage occurred. The range-through method was used (Labandeira and Sepkoski 1993), in which the first occurrence datum and last occurrence datum defined the continuous presence of the lineage, whether or not it has been recorded in intervening stages. Occurrence data for the Jurassic and Cretaceous periods were plotted at the level of the geologic stage, whereas Paleogene stage-level data were amalgamated at the more inclusive level of the geologic epoch. The most recent, internationally approved, standardized geochronology was used (Gradstein et al. 2012).

A second set of guidelines circumscribed the early angiosperm fossil record. Background information for the Jurassic through Paleogene record of land plants originated from several sources, including the mutually consistent and occasionally rich palynological, mesofossil, and macrofossil records (Friis et al. 2011). Of relevance to data collection is the origin of angiosperms during the early Cretaceous Period, consistent with a wide variety of paleobotanical and plant-morphological evidence (Crane et al. 1995), and increasingly with molecular evidence (Bell et al. 2010; Magallón 2010). The origin of angiosperms is taken as no earlier than the mid-Hauterivian stage at ca. 135 Ma (Friis et al. 2011). The subsequent, primary diversification interval of angiosperms occurred during the 35 million year interval

from the Barremian–Aptian stage boundary to the Turonian–Coniacian stage boundaries (Hughes 1994).

A third group of procedures were employed to establish the presence of fossil insect lineages. Several compendia were consulted to determine occurrence data for fossil insect lineages (Dmitriev and Zherikhin 1988; Rasnitsyn 1988; Carpenter 1992; Ross and Jarzembowski 1993; Labandeira 1994; Evenhuis 1994; Rasnitsyn and Quicke 2002; Grimaldi and Engel 2005; Sohn et al. 2012), buttressed by updates from recent taxonomic insect literature and the online Paleobiology Database (PBDB, 2014), accessed through the Fossil Works portal. As many of the earlier compendia had spurious occurrences, it was essential to consult considerably more modern sources to rectify synonymies, delete unvetted data, add new occurrences, and provide more current time-range extensions or contractions. After these filters were used, the culled dataset consisted of 280 family-level fossil insect lineages. The family was the focal taxonomic rank of interest. Alternative, more modern, classifications occasionally demote families to subfamily rank, a consequence that was taken into account in constructing Fig. 13.2. The insect lineage dataset consisted of 36.8 % Hemiptera, by far the most represented group; ca. 14 % each of Coleoptera and Lepidoptera; ca. 10 % each for Diptera and Hymenoptera; ca. 4–6 % each for Orthoptera and Thysanoptera, and 1–1.5 % each for the least abundant lineages of the Phasmatodea, Neuroptera, Trichoptera, and Mecoptera (Table 13.1, Fig. 13.1).

13.2.3 Establishing Feeding Guild, Functional Feeding Group and Plant–Host Assignments

Eight criteria were used to establish plant–host assignments of herbivory, pollination, xylophagy, and mimicry. These criteria can be divided into habitat-related ecological features and insect-specific morphological attributes. For broad-scale ecological features, the first consideration consists of broad, host–plant affiliations and related ecological attributes of modern descendant taxa, particularly if significant agricultural, entomological, or botanical information is available (Labandeira 1998). This process is taxonomic uniformitarianism (Dodd and Stanton 1990), and assumes that no or minimal host–plant shifts have occurred since the earliest fossil occurrence of the insect lineage in question. A second criterion involves the taxonomic spectrum of herbivorized plants of the flora in which an insect taxon co-occurs. Obviously, the host preferences of an insect in a preangiospermous flora can be safely attributed to a cryptogam, fern or gymnosperm. Conversely, an insect occurring in a diverse flora and consisting only of angiosperms can reasonably be associated with an angiosperm host. A third criterion involves specification of a particular damage type (Labandeira et al. 2007b) that could be attributed to a certain, family-level taxon. An example is the assignment of distinctive leaf mines occurring on angiosperm leaves of a sycamore host species (Platanaceae) from the early Paleocene of Montana, United States, to the dipteran family Agromyzidae (Winkler et al. 2010).

Four additional criteria indicate that host affiliations may be based on morphological features. The fourth criterion is the mouthpart structure of a representative insect taxon from the group in question (Labandeira 1997), which in some instances can be linked to particular types of herbivore damage, pollinator access, or wood boring in the same flora. An example would be the distinctive and specialized phytophagous mouthparts of weevils from the Yixian Formation in northeastern China (Davis et al. 2013), that also would imply gymnospermous plant hosts. Fifth, is presence of gut contents consisting of plant material (Rasnitsyn and Krassilov 2000) or pollen (Krassilov et al. 2007), which provides direct evidence of host affiliations of the insect consumer. Sixth, for pollinator assignment to plant hosts, certain features can be important, such as pollen plastered or attached to the mouthparts, or ventral aspect of the head capsule the associated insect with specialized, pollen-gathering structures such as bee corbiculae (Engel 2000) or thrips ring setae (Peñalver et al. 2012). The seventh criterion, also applicable to pollinators, is the presence of particular plant features that would indicate pollination (Labandeira et al. 2007a). For gymnosperm pollinators of the mid-Mesozoic, probed structures such as integumental tubes, deep funnels, and channels in ovulate organs were used by long-proboscid insects to access nectar-like pollination drops (Ren et al. 2009).

Last, in the case of mimicry, occasionally plant foliage shares an uncanny, detailed resemblance (the models) to particular co-occurring insect species (the mimics). Examples include strong resemblance of wings from one neuropteran species to a particular fern pinnule (Wang et al. 2010); or the entire body of another neuropteran species to a particular ginkgophyte leaf (Wang et al. 2012b).

13.2.4 Rationale for Understanding Gymnosperm-to-Angiosperm Host Transitions

The initial phase of angiosperm diversification established all major angiosperm lineages during a 35 million-year-long interval that encompassed the four mid-Cretaceous stages of the Aptian, Albian, Cenomanian, and Turonian. It would have been during this time interval that many insect lineages associated with gymnosperm hosts but known to have angiosperm-dominant associations in the more recent part of the geologic record would have shifted to angiosperm hosts (Tables 13.1, 13.2, and 13.3). Given that the angiosperm radiation is represented by four geologic stages during which the shift occurred, transfer ratios were allocated to each of the four constituent stages to represent a linear, monotonic shift from gymnosperm to angiosperm hosts. For the Aptian stage, 25 % of 60 insect lineages were transferred from gymnosperm → angiosperm hosts (column D of Tables 13.2 and 13.3, in bold lettering) to angiosperm-only hosts (column E of Tables 13.2 and 13.3); analogous values for the Albian stage were 50 % of 56 families; for the Cenomanian, 75 % of 63 families; and for the Turonian, 100 % of 64 families, after

Table 13.3 The Mid-Mesozoic transition from gymnosperm- to angiosperm-dominated host plants[a]

Period and time interval[c]	Geologic stage or epoch[d]	Dominant host-plant preferences of phytophagous lineages[b]					Number of families and their transfer ratio from column *D* to column *E* during the four stages of the angiosperm radiation[e]
		A Cryptogam and fern hosts only	B Cryptogam → angiosperm transitions	C Gymnosperm hosts only	D Gymnosperm → angiosperm host transitions	E Angiosperm hosts only	
Paleogene (23–66 Ma)	Oligocene	2	0	12	0	179	
	Eocene	2	0	13	0	171	
	Paleocene	1	0	11	0	121	
Cretaceous (66–145 Ma)	Maastrichtian	1	0	10	0	111	
	Campanian	1	0	13	0	111	
	Santonian	1	0	16	0	110	
	Coniacian	1	0	14	0	104	
	Turonian	**1**	**0**	**14**	**0**	**104**	**100 % of 64 families**
	Cenomanian	**1**	**1**	**14**	**16**	**87**	**75 % of 63 families**
	Albian	**1**	**3**	**17**	**27**	**58**	**50 % of 56 families**
	Aptian	**2**	**3**	**20**	**45**	**33**	**25 % of 60 families**
	Barremian	3	1	37	58	8	
	Hauterivian	4	1	39	57	2	
	Valanginian	4	1	38	55	0	
	Berriasian	4	1	40	55	0	

(continued)

Table 13.3 (continued)

Period and time interval[c]	Geologic stage or epoch[d]	Dominant host-plant preferences of phytophagous lineages[b]					Number of families and their transfer ratio from column *D* to column *E* during the four stages of the angiosperm radiation[e]
		A Cryptogam and fern hosts only	B Cryptogam → angiosperm transitions	C Gymnosperm hosts only	D Gymnosperm → angiosperm host transitions	E Angiosperm hosts only	
Jurassic (145–201 Ma)	Tithonian	3	1	41	44	0	
	Kimmeridgian	2	1	43	42	0	
	Oxfordian	4	0	44	39	0	
	Callovian	2	0	49	26	0	
	Bathonian	1	0	38	14	0	
	Bajocian	1	0	32	14	0	
	Aalenian	1	0	29	14	0	
	Toarcian	1	0	28	14	0	
	Pliensbachian	1	0	27	10	0	
	Sinemurian	1	0	27	9	0	
	Hettangian	1	0	26	7	0	

[a] Data are taken from Fig. 13.1

[b] There are 280 phytophagous insect families used in this table (Fig. 13.1)

[c] Geologic time periods are in bold; age dates and all geochronologic time intervals are from Gradstein et al. (2012)

[d] The Paleogene period is divided into epochs whereas the Cretaceous and Jurassic periods are divided into stages. The four stages of the angiosperm radiation, the Aptian through Turonian stages inclusive, are highlighted in bold

[e] These ratios represent the percentage of major gymnosperm-to-angiosperm host–plant transitions occurring during the angiosperm radiation. See text for details

which all families with gymnosperm-to-angiosperm host transitions were tabulated in the angiosperm-only host column. These transfers are independent of the gymnosperm-only host column which retained dominantly gymnosperm hosts but never acquired dominantly angiosperm hosts.

In a manner parallel to that of the gymnosperms detailed above, insect lineages associated with cryptogams and ferns were evaluated during the angiosperm radiation (column A of Tables 13.2 and 13.3). Similarly, lineages were assessed that possessed dominant cryptogam or fern associations which shifted to angiosperm-dominant associations during the angiosperm radiation (column B of Tables 13.2 and 13.3).

13.3 Results

Figure 13.2 depicts range-through occurrences of 280 vertically arrayed, family-level lineages that represent eleven plant-associated insect orders along a Jurassic through Paleogene time series. The insect lineages are characterized by the dominant feeding modes of herbivore, pollinator, herbivore–pollinator, pollinator–mimic, and xylophage, and whether their dominant hosts are cryptogams/ferns (purple), gymnosperms (darker hue), or angiosperms (lighter hue), as indicated in the legend insets. Major gymnosperm-to-angiosperm host–plant transitions during the angiosperm radiation (gray vertical column) are indicated. The data in Fig. 13.2 are restated in Table 13.3, which is a geochronologic stage-by-stage summary of the raw data in Table 13.1. Summary Fig. 13.3 details the trivariate relationship between (i) the diversity of fossil insect families in the vertical axis), (ii) their major plant–host associations of cryptogams/ferns (purple), gymnosperms (dark green), and angiosperms (light green) in the field of the figure, and (iii) stage-level geologic time in the horizontal axis. While not expressed graphically, the functional-feeding-group data in Table 13.1 is presented in summary form in the middle columns of Table 13.2. These data provide a qualitative description of functional feeding strategies for each insect lineage that are not apparent from their role in a dominant feeding guild or from their host–plant associations.

13.3.1 Plant-Feeding Features of Jurassic to Paleogene Insect Lineages

The dataset of 280 plant-associated insect families are categorized by order and partitioned into three feeding-related ecological attributes (Table 13.2). The first feeding attribute is the dominant feeding guild, the second is the major functional feeding group, and the third is the dominant plant hosts and their transitions. For the dominant feeding guild, the most frequently encountered category are herbivores (56.4 % of all occurrences), then herbivore–pollinators (19.6 %), pollinators

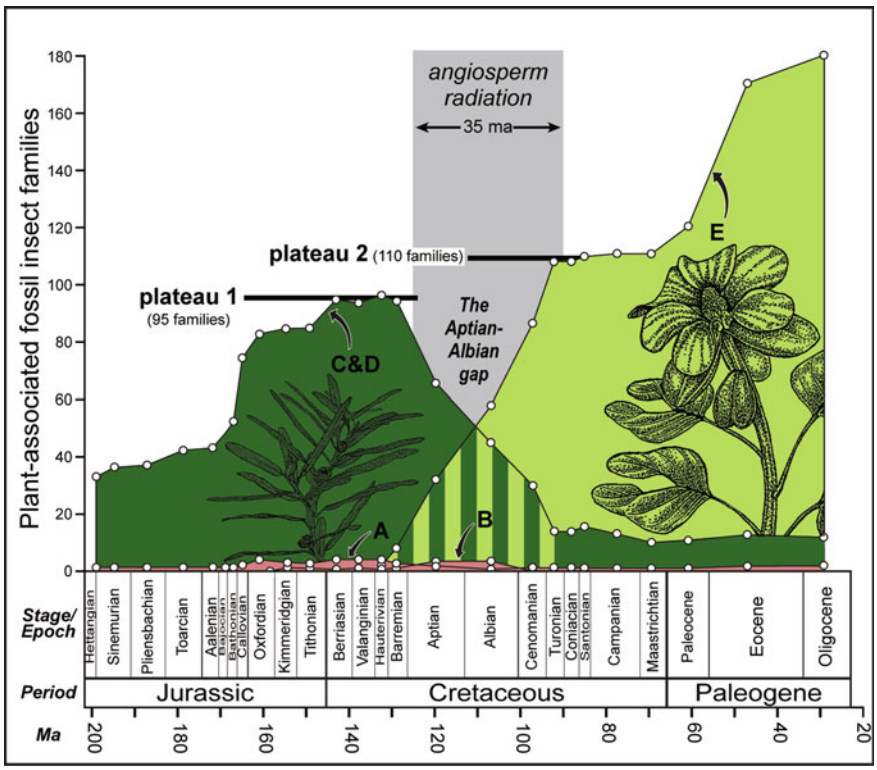


Fig. 13.3 Plot of the major plant hosts (*field of view*) associated with plant-associated fossil insect families (*vertical axis*) versus geologic time (*horizontal axis*). Data are derived from Table 13.1, summarized in Table 13.3. The *purple* color indicates cryptogam and fern hosts; *dark green* indicates gymnosperm hosts; *light green* indicates angiosperm hosts. The *vertical column* indicates the interval of time represented by the initial angiosperm radiation

(17.5 %), xylophages (5.0 %), and pollinator–mimics (1.4 %). Of these five categories, herbivory dominates all other interactions in the dataset.

For the functional feeding group (Table 13.2), the most frequently encountered mode is piercing and sucking (34.6 % of all occurrences), then external feeding (28.2 %), surface fluid feeding (14.3 %), galling (6.4 %), wood boring (5.7 %), leaf mining (4.3 %), palynivory (3.6 %) and seed predation (2.9 %). These proportions indicate that external (ectophagous) feeding predominates for ca. four-fifths of the occurrences whereas internal (endophagous) feeding contributes to only one-fifth of the data.

The third attribute is the identity of the dominant plant hosts and the amount of host switching among the dominant insect families. Those lineages with angiosperms as the dominant host represent 42.9 % of all families, whereas those with gymnosperms as the dominant hosts consisted of 29.3 %. Lineages hosting gymnosperm hosts and existing prior to the advent of angiosperms but later shifting to angiosperm-dominant

hosts provided 24.3 % of families. Cryptogams and ferns played a minor role as major hosts of insect lineages, consisting of 1.4 % of all cryptogam/fern-only occurrences and similarly 1.1 % of all cryptogam/fern lineages transitioning onto angiosperms.

13.3.2 Gymnosperm Versus Angiosperm Host Use Before, During, and After the Angiosperm Radiation

Data from Table 13.3 are plotted in Fig. 13.3. Shown in purple for Fig. 13.3 are insect families that retained their dominant cryptogam/fern hosts to the end of the Paleogene Period (trajectory A), recorded from column A of Table 13.3; and those that shifted dominantly to angiosperms during the angiosperm radiation (trajectory B), recorded from column B of Table 13.3. Likewise, shown in dark green are insect families that have kept their dominantly gymnosperm hosts, recorded from column C of Table 13.3, to which are added those insect lineages that transitioned from earlier gymnosperm-dominant hosts to angiosperm-dominant hosts after the angiosperm radiation, recorded from column D of Table 13.3 in bold lettering (See Sect. 13.2.4 for details). Insect family-level diversity with gymnosperm hosts thus represent the summation of columns C and D in Table 13.3, plotted as trajectory C + D in Fig. 13.3. Insect families with dominantly angiosperm hosts and originating during or after the angiosperm radiation are shown in light green and provide the most sustained increase of a host-affiliated insect group (trajectory E).

Two derivative features involving insect families with particular plant–host affiliations are depicted in Fig. 13.3. First, insect families with gymnosperm-dominant plant hosts, shown in trajectory C + D, form a distinct diversity plateau of ca. 95 families during the 20 million year-long Berriasian to Barremian interval, perhaps extending back in time to a decreased level of ca. 85 families to the Oxfordian stage another 20 million years earlier. After the Barremian stage, and the angiosperm radiation, insect lineages with gymnosperm hosts decrease linearly and monotonically to a flat diversity level of 10–14 insect families. By contrast, insect families with angiosperm-dominant hosts of trajectory E increase linearly and monotonically commencing at the angiosperm radiation, and reaching a sustained plateau of 110 families for the ca. 20 million years of the Turonian through Maastrichtian stages. Thereafter, insect families with angiosperm-dominant hosts increase dramatically into the late Paleogene.

Other than these two diversity plateaus, Fig. 13.3 illustrates a distinctive gap between the trajectories of C + D and E. Before and after the crossover between the gymnosperm-dominant and angiosperm-dominant family diversity curves of insects, there is a collective diversity minimum, the Aptian–Albian gap. The Aptian–Albian gap spans the angiosperm radiation and represents a significant decrease of 45 % from the earlier gymnosperm plateau of 95 families and 53 % of the later angiosperm plateau of 110 families.

13.4 Discussion

Three broader aspects of these findings deserve an extended mention. An obvious issue is to what extent does the data presented here explain the presumed "counterintuitive" result reported in Labandeira and Sepkoski (1993) that there was no increase in insect diversity during the formative interval of initial angiosperm diversification (Fig. 13.1)? Secondly, what is the meaning of the earlier gymnosperm and later angiosperm plateaus that bracket the angiosperm radiation, and do they have any relationship to the intervening Aptian–Albian gap (Fig. 13.3)? Last, is there a broader message about attempting to understand the role of fossil insect diversity vis-à-vis the angiosperm radiation, and vice versa, by using multiple approaches of investigation (Fig. 13.4).

13.4.1 Reasons for the Mid-Mesozoic Constancy of Insect Family-Level Diversity

There are several, independent explanations that could explain the relative stasis of family-level insect diversity during the angiosperm radiation. One reason, based on evidence from this report, is that an expectation of elevated insect diversification during the angiosperm radiation that would range from diffuse to intimate coevolution (Friis et al. 2011), needs to be balanced by evidence indicating equally high associational diversity between insects and gymnosperms prior to the angiosperm radiation. Given recent developments in understanding the associational diversity between gymnosperms and insects prior to and during the angiosperm radiation (e.g., Ratzel et al. 2001; Ren 1998; Ren et al. 2009; Labandeira 2010; Wang et al. 2012b; Peñalver et al. 2012; Ding et al. 2014), it is highly likely that gymnosperm–insect interactions preceding the angiosperm radiation were almost or just as diverse as angiosperm interactions that followed the event.

A second reason involves the Mesozoic Lacustrine Revolution, which evidently changed food-web structure of lotic and lentic ecosystems during the late Jurassic to early Cretaceous (Buatois et al. 2015). The environmental context of this transformation involves the change from detritivore based, typically hypotrophically stratified water bodies (Zherikhin et al. 1999), to herbivore dominated, typically pseudoligotrophically stratified water bodies (Sinitshenkova 2002). This physiochemical and biological turnover in aquatic ecologic structure occurred during the mid-Cretaceous and is synchronous with an aquatic insect extinction event (Sukatsheva 1991; Buatois et al. 2014). Approximately 20 family-level insect lineages became extinct at the Mesozoic Lacustrine Revolution (Buatois et al. 2014), supplemented by an additional 30 % of the plant-associated insect families during the same time interval.

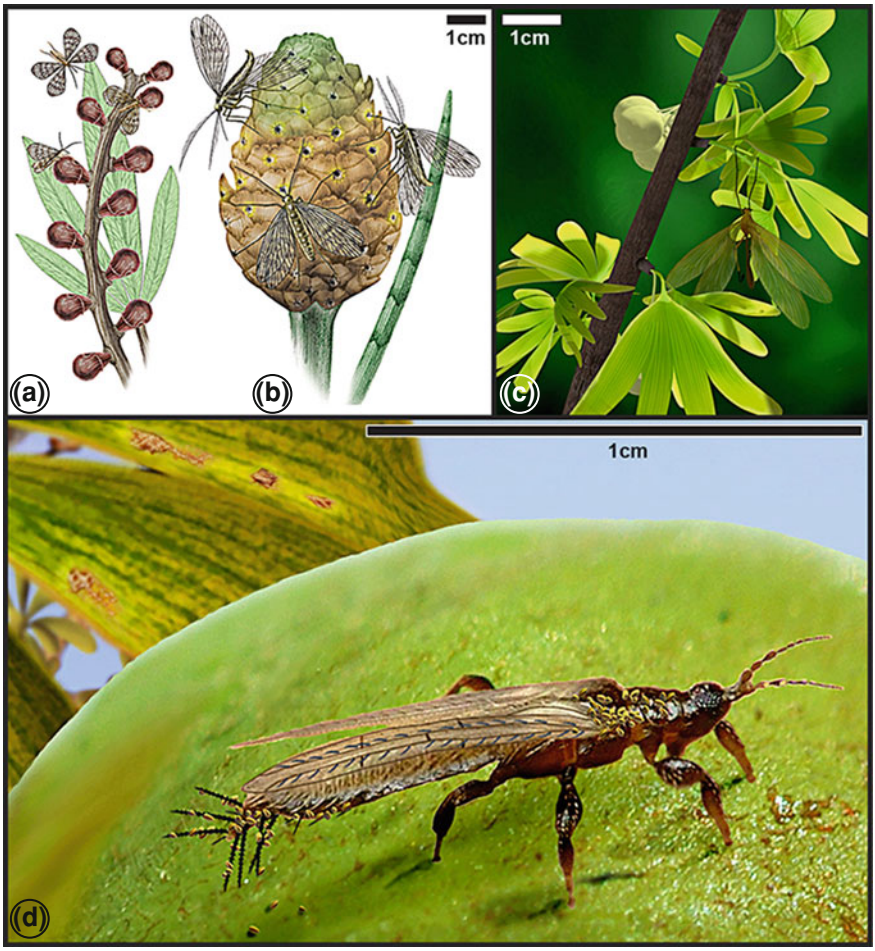


Fig. 13.4 Three mecopteran and a thysanopteran insect association with mid-Mesozoic gymnosperms. **a** The mecopteran long-proboscid pollinator, *Lichnomesopsyche gloriae* (Mesopsychidae, entry 224 of Table 13.1 and Fig. 13.2), with proboscis entering an integumental channel in the ovulate organ of *Caytonia sewardi* (Caytoniaceae) from the Callovian of Inner Mongolia, China. **b** Another mecopteran long-proboscid pollinator, *Vitimopsyche kozlovi* (Mesopsychidae, also entry 224 of Table 13.1 and Fig. 13.2), bearing *Classopollis* pollen and probing the ovulate organ catchment funnel of *Alvinia bohemica* (Cheirolepidiaceae) from the Barremian of Liaoning, China. **c** The mecopteran leaf mimic, *Juracimbrophlebia ginkgofolia* (Cimbrophlebiidae, entry 223 of Table 13.1 and Fig. 13.2), resembling a multilobed *Ginkgoites* leaf of *Yimaia capituliformis* (Ginkgoaceae) from the Callovian of Inner Mongolia, China. **d** The thysanopteran punch-and-suck pollinator, *Gymnopollisthrips minor* (Melanothripidae, entry 25 of Table 13.1 and Fig. 13.2), with *Cycadopites* sp. pollen grains on the ovulate organ of *Nehvezdyella bipartita* (Nehvezdyellaceae) from the Aptian of Spain. Drawings reprinted with permission: (**a**) and (**b**) courtesy of Mary Parrish, N.M.N.H. Department of Paleobiology in Washington, DC, USA; (**c**) courtesy of Wang Chen, C.N.U. College of Life Sciences in Beijing, China; and (**d**) courtesy of Enrique Peñalver, Instituto Geológico y Minero de España in Madrid, Spain)

A third cause for the constancy of diversity involves the parasitoid diversification of especially Hymenoptera, and to a lesser extent Diptera, during the Jurassic and Early Cretaceous (Rasnitsyn 1980). This major radiation of major, high-ranked lineages (Labandeira 2002) had a major effect not only on top to down regulation of herbivores in terrestrial webs, but significantly increased Jurassic and Early Cretaceous insect diversity that are captured in global compendia (Table 13.1, footnote 2) and in derivative plant–insect studies (Labandeira and Sepkoski 1993; Jarzembowski and Ross 1996; Alekseev et al. 2001). The inclusion of parasitoid insect families prior to the angiosperm radiation would have the effect of increasing insect diversity and balancing diversity levels after the angiosperm radiation.

Last, there is considerable evidence from modern molecularly based phylogenetic analyses that some plant-associated insect lineages diversified preceding the angiosperm radiation. Evidence for this comes from the major hyperdiverse clades such as the Hemiptera (Moran et al. 2005; Cocroft et al. 2008; Wang et al. 2012a), Hymenoptera (Rasnitsyn 1980; Davis et al. 2010), Coleoptera (Farrell 1998; McKenna et al. 2009; Wang et al. 2013), Diptera (Ren 1998; Labandeira 2005), and Lepidoptera (Imada et al. 2011; Zhang et al. 2013). The deeper extensions suggest that these lineages were diverse and actively consuming live tissues of cryptogams, ferns, and gymnosperms from millions to a few tens of millions of years before the initial angiosperm diversification interval.

Contributions of family-level insect taxa from these four data sources would provide relative stasis in family-level insect diversity throughout the late Jurassic and into the early Paleogene (Fig. 13.1). However, of these data sources, plant–host-associated families of insects likely were most important.

13.4.2 Host Switching, Diversity Plateaus and the Aptian–Albian Diversity Low

The pattern of gymnosperm-to-angiosperm host–plant dominance throughout the 100 million year interval from the Callovian (166 Ma) to the K-Pg extinction event at the end of the Cretaceous (66 Ma) potentially reveals the family-level insect dynamics associated with this shift (Fig. 13.3). An upper limit of 95 families was reached for insect lineages whose hosts were dominantly to exclusively gymnosperms, supplemented by a minor level of cryptogam and fern associations. This plateau disappeared at the beginning of the angiosperm radiation, as major older insect lineages with gymnosperm (and cryptogam and fern) hosts shifted onto new angiosperm hosts (Peñalver et al. 2012; Ding et al. 2014), or became extinct, and as new major insect lineages initially hosting angiosperms increased dramatically in diversity (Labandeira et al. 1994; Hartkopf-Fröder et al. 2011). By the Santonian stage (85 Ma) of the Late Cretaceous, an upper limit of 110 families was established, which remained until a dramatic diversity increase following the K-Pg event.

Separating the earlier plateau of gymnosperm-dominated families from subsequent and somewhat more elevated plateau of angiosperm-dominated families is

the Aptian–Albian gap, which represents an interval of time characterized by transition from gymnosperm to angiosperm hosts. The probable cause of this gap is time lags that occur between when a food resource is available and when it becomes herbivorized. Geochronologic lags have been demonstrated at time intervals such as the appearance of vascular plant tissues in the earlier Devonian, and when they are later herbivorized during the mid-Paleozoic (Labandeira 2007); additionally, time delays occur in the colonization of eudicot plant hosts by lepidopteran leaf-mining genera during the late Cretaceous and Cenozoic (Lopez-Vaamonde et al. 2006). This downturn in plant-associated insect diversity is evident in the coarse-grained epoch-level analysis of Jarzembowski and Ross (1993) and the fine-grained stage-level analyses of Alekseev et al. (2001), although the gap appears phase-shifted toward the late Cretaceous by a stage or two. An Aptian–Albian minimum also may be present in aquatic insect lineages (Buatois et al. 2014), but no relevant analysis of family-level aquatic insect diversity has been made for this interval.

The presence of two, successive upper bounds for insect lineages with gymnosperm- and angiosperm-dominant hosts separated by a diversity minimum (Fig. 13.3), suggests a global event and major plant–host replacement during this 35 million-year interval. Such an event would represent a significant scaling up from considerably more spatiotemporally and taxonomically confined host shifts illustrated between plants and insects from a variety of modern habitats (Pellmyr and Seagraves 2003; Cocroft et al. 2008). One particular system is nymphaline butterflies (fritillaries) and their angiosperm lamialean (mints and relatives) and asteraceous (daisies and relatives) hosts. One study (Nylin and Wahlberg 2008) indicates that host shifts of angiosperm food plants by fritillary butterflies likely were associated with a previous extensive period of polyphagy, wherein multiple, unrelated, colonized plant lineages were replaced by a major shift to a novel host unrelated to the previous spectrum of consumed plants. This host shift was accelerated by a major plant extinction event at the K-Pg boundary, which restricted the range of new potential hosts available to certain fritillaries (Nylin and Wahlberg 2008). Significant range expansions of post-event surviving fritillary taxa may have enhanced the probability of new host shifts (Weingartner et al. 2006). Such an event, between fritillaries and their dicot hosts, when multiplied and writ large geographically, could provide a model for understanding the extensive, global gymnosperm-to-angiosperm shift of many insect lineages during the mid-Cretaceous.

13.4.3 Questions of the Fossil Record That Only Can Be Answered by Multidisciplinary Data

Back in 1993, the study by Labandeira and Sepkoski was purely an exploratory venture toward understanding the insect fossil record. One of the patterns noted by some and engendering significant negative animus was the pattern of long-term stasis of insect families that encompassed the considerable stretch of time from the Late Jurassic to the early Paleogene, notably including the angiosperm radiation.

This result provided an opportunity to subsequently pursue alternative research to test the conclusion of Labandeira and Sepkoski (1993) that the angiosperm radiation had no effect on family-level insect diversity. One approach assessed features of suspect insect pollinators in preangiospermous floras to determine whether gymnosperms were being actively pollinated (Fig. 13.4a, b, d) (Labandeira et al. 2007a; Ren et al. 2009; Labandeira 2010; Peñalver et al. 2012). A second tack involved the presence, extent, and type of mimicry in preangiospermous biotas (Fig. 13.4c) (Wang et al. 2010, 2012b). A third opportunity allowed investigation of early angiospermous and older gymnospermous Mesozoic floras to establish quantitative levels of herbivory diversity and intensity before and after the angiosperm radiation (Ding et al. 2014). A fourth procedure is the examination of ecosystem food-web structure before (currently no data) and after (Dunne et al. 2014) the angiosperm radiation. And last, modern molecularly based evolutionary biology studies indicate that significant radiations of plant-associated insect lineages occurred earlier than the angiosperm radiation (Farrell 1998; Davis et al. 2010). These approaches suggest that the ecological and evolutionary biological infrastructure of insect lineages associated with gymnosperm hosts was a 40 million-year-long feature of Late Jurassic to mid-early Cretaceous terrestrial habitats. This process was rivaled by insect lineages occurring on diverse angiosperm hosts that persisted for a 20 million-year interval during the Late Cretaceous.

13.5 Conclusions

This report should be seen as a first attempt in addressing the reciprocal roles of insects and angiosperms during the initial radiation of angiosperms. Although there are three major conclusions derived from the data presented in this report, additional analyses with more improved, taxonomically resolved and geochronologically constrained data would go far to ferret out further these patterns.

1. *The angiosperm radiation.* Herbivory is one of the fundamental attributes of insects, and the host plants consumed by insects are resources that have an important fossil record. One of the major episodes in the evolution of insect herbivory is the transition from gymnosperm- to angiosperm-dominant hosts during the initial diversification of angiosperms 125 to 90 million years ago. Exploring the patterns and evolutionary and ecological mechanisms responsible for this global taxonomic shift in consumer resources is a goal of this report.
2. *The pattern.* Evidence indicates that plant-associated insect families that hosted gymnosperms prior to the angiosperm radiation consisted of a sustained peak of ca. 95 families for 40 million years. This was followed by a switchover by insect lineages that acquired angiosperm hosts, eventually reaching a level of ca. 110 lineages for a 20 million-year interval during the Late Cretaceous. After this stasis in diversity, there was a rapid increase in angiosperm-hosted insect lineages well into the Paleogene Period. Thus, a major gap occurred

during these two diversity maxima levels present on both sides of the angiosperm radiation, attributable to turnover in the plant–host preferences of insect lineages and time-lag effects resulting in the shift from gymnosperm to angiosperm hosts. Notably, the plateau established by earlier insect lineages with gymnosperm hosts was 86.4 % that of the later insect lineages with angiosperm hosts.

3. *Implications.* The pattern and inferred processes outlined here indicates that modern insect lineages retain only a very minor legacy of their former Middle Jurassic to mid-Cretaceous gymnosperm hosts. By contrast, insect lineages with dominantly gymnosperm hosts during the preangiospermous Late Jurassic to Early Cretaceous rivaled in diversity insect lineages with dominantly angiosperm hosts after the mid-Cretaceous angiosperm radiation and throughout the Late Cretaceous. The ecology of interactions between these older insect lineages and their dominantly gymnosperm hosts needs to be explored further to establish an entrée into this earlier world devoid of angiosperms.

Acknowledgments Thanks go to Pierre Pontarotti for inviting CCL to attend the Seventeenth Evolutionary Biology meeting in Marseille France. Finnegan Marsh adroitly crafted the figures. We are grateful for the Missouri Botanical Garden (St. Louis, Missouri), Wang Chen (Capital Normal University, Beijing), and Enrique Peñalver (Museo Geominero, Madrid) for use of images in Fig. 13.4. An anonymous reviewer improved the manuscript. Use of the online Paleobiology Data Base (PBDB) is acknowledged. This is contribution 263 to the Evolution of Terrestrial Ecosystems consortium at the National Museum of Natural History, in Washington, D.C.

References

Alekseev AC, Dmitriev VY, Ponomarenko AG (2001) The evolution of taxonomic diversity. Geos, Moscow

Barreda VD, Cúneo NR, Wilf P, Currano ED, Scasso RA, Brinkhuis H (2012) Cretaceous/Paleogene floral turnover in Patagonia: drop in diversity, low extinction, and a *Classopollis* spike. PLoS ONE 7(12):e52455

Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-revisited. Am J Bot 97:1296–1303

Buatois LA, Labandeira CC, Cohen AC, Mángano G, Voigt S (2015) The fossil history of continental aquatic taxa and the Mesozoic lacustrine revolution. In: Buatois LA, Mángano G (eds) The trace-fossil record of major evolutionary events, in review. Springer, Berlin (in review)

Carpenter FM (1992) Superclass Hexapoda. In: Moore RC, Kaesler RL, Brosius E, Keim J, Priesner J (eds) Treatise on invertebrate paleontology, part R Arthropoda 4, vol 3 and 4. Geological Society of America, Boulder, and University of Kansas, Lawrence

Cocroft RB, Rodriguez RL, Hunt RE (2008) Host shifts, the evolution of communication, and speciation in the *Enchenopa binotata* species complex of treehoppers. In: Tillmon KJ (ed) Specialization, speciation, and radiation: the evolutionary biology of herbivorous insects. University of California Press, Berkeley, pp 88–100

Crane PR (1987) Vegetational consequences of the angiosperm diversification. In: Friis EM, Chaloner WG, Crane PR (eds) The origins of angiosperms and their biological consequences. Cambridge, New York, p 107–144

Crane PR, Friis EM, Pedersen KR (1995) The origin and early diversification of angiosperms. Nature 374:27–33
Currano ED, Labandeira CC, Wilf P (2009) Fossilized insect folivory tracks temperature for six million years. Ecol Mon 80:547–567
Davis RB, Baldauf SL, Mayhew PJ (2010) The origins of species richness in the Hymenoptera: insights from a family-level supertree. BMC Evol Biol 10:109
Davis SR, Engel MS, Legalov A, Ren D (2013) Weevils of the Yixian Formation, China (Coleoptera: Curculionoidea): phylogenetic considerations and comparison with other Mesozoic faunas. Syst Entomol 11:399–429
Ding Q, Labandeira CC, Ren D (2014) Distinctive insect leaf mines on *Liaoningocladus boii* (Coniferales) from the Early Cretaceous Yixian Formation of northeastern China. Arthro Syst Phylo (in review)
Dmitriev VJ, Zherikhin VV (1988) Changes in the diversity of insect families from data of first and last occurrences. In: Ponomarenko AG (ed) The Mesozoic-Cenozoic crisis in the evolution of insects. Nauka, Moscow, pp 208–215 (in Russian)
Dodd JR, Stanton (1990) Paleoecology: concepts and applications, 2nd edn. Wiley, New York
Dolling WR (1991) The Hemiptera. Oxford, New York
Dunne JA, Labandeira CC, Williams RJ (2014) Highly resolved middle Eocene food webs show early development of modern trophic structure after the end-Cretaceous extinction. Proc Roy Soc B 281. http://dx.doi.org/10.1098/rspb.2013.3280
Engel MS (2000) A new interpretation of the oldest fossil bee (Hymenoptera: Apidae). Am Mus Novit 3296:1–11
Evenhuis NL (1994) Catalogue of the fossil flies of the world (Insecta: Diptera). Backhuys, Leiden
Farrell BD (1998) "Inordinate fondness" explained: why are there so many beetles? Science 281:555–559
Friis EM, Pedersen KR, Crane PR (2010) Diversity in obscurity: fossil flowers and the early history of angiosperms. Phil Trans Roy Soc B 365:369–382
Friis EM, Crane PR, Pedersen KR (2011) Early flowers and angiosperm evolution. Cambridge, Cambridge
Gauld I, Bolton B (eds) (1988) The Hymenoptera. Oxford, New York
Goulet H, Huber JT (eds) (1993) Hymenoptera of the world: an identification guide to families. Agriculture Canada, Ottawa
Gradstein FM, Ogg JG, Schmitz MD, Ogg G (2012) The geologic time scale 2012. Elsevier, Boston
Grimaldi D, Engel MS (2005) Evolution of the insects. Cambridge, New York
Harris TM (1942) *Wonnacottia*, a new Bennettitalean microsporophyll. Ann Bot 6:577–592
Hartkopf-Fröder C, Rust J, Wappler T, Friis EM, Viehofen A (2011) Mid-Cretaceous charred flowers reveal direct observation of arthropod feeding strategies. Biol Lett 8:295–298
Hill RS, Carpenter RJ (1999) *Ginkgo* leaves from Paleogene sediments in Tasmania. Austral J Bot 47:717–724
Hughes N (1994) The enigma of angiosperm origins. Cambridge University Press, Cambridge
Imada Y, Kawakita A, Kato M (2011) Allopatric distribution and diversification without niche shift in a bryophyte-feeding basal moth lineage (Lepidoptera: Micropterigidae). Proc Roy Soc B 278:3026–3033
Jarzembowski EA, Ross AJ (1993) Time flies: the geological record of insects. Geol Today 9:218–223
Jarzembowski EA, Ross AJ (1996) Insect origination and extinction in the Phanerozoic. In: Hart MB (ed) Biotic recovery from mass extinction events. Geol Soc Spec Publ 102:65–78
Krassilov VA, Rasnitsyn AP, Afonin SA (2007) Pollen eaters and pollen morphology: co-evolution through the Permian and Mesozoic. Afr Invert 48:3–11
Labandeira CC (1994) A compendium of fossil insect families. Milwaukee Publ Mus Contrib Biol Geol 88:1–71

Labandeira CC (1997) Insect mouthparts: ascertaining the paleobiology of insect feeding strategies. Annu Rev Ecol Syst 28:153–193
Labandeira CC (1998) Early history of arthropod and vascular plant associations. Annu Rev Earth Planet Sci 26:329–377
Labandeira CC (2002) The paleobiology of predators, parasitoids, and parasites: accommodation and death in the fossil record of terrestrial invertebrates. In: Kowalewski M, Kelley PH (eds) The fossil record of predation. Paleontol Soc Pap 8:211–250
Labandeira CC (2005) Fossil history and evolutionary ecology of Diptera and their associations with plants. In: Yeates DK, Wiegmann BM (eds) The evolutionary biology of flies. Columbia, New York, pp 217–273
Labandeira CC (2006) Silurian to Triassic plant and insect clades and their associations: new data, a review, and interpretations. Arth Syst Phylo 64:53–94
Labandeira CC (2007) The origin of herbivory on land: the initial pattern of live tissue consumption by arthropods. Ins Sci 14:259–274
Labandeira CC (2010) The pollination of mid Mesozoic seed plants and the early history of long-proboscid insects. Ann Missouri Bot Gard 97:469–513
Labandeira CC (2013) A paleobiological perspective on plant–insect interactions. Curr Opin Pl Biol 16:414–421
Labandeira CC, Allen EM (2007) Minimal insect herbivory for the lower Permian coprolite bone bed site of north-central Texas, USA, and comparison to other late Paleozoic floras. Palaeogeogr Palaeoclimatol Palaeoecol 247:197–219
Labandeira CC, Dilcher DL, Davis DR, Wagner DL (1994) Ninety-seven million years of angiosperm-insect association: paleobiological insights into the meaning of coevolution. Proc Natl Acad Sci USA 91:12278–12282
Labandeira CC, Kvaček J, Mostovski MB (2007a) Pollination fluids, pollen, and insect pollination of Mesozoic gymnosperms. Taxon 56:663–695
Labandeira CC, Johnson KR, Wilf P (2002) Impact of the terminal Cretaceous event on plant–insect associations. Proc Natl Acad Sci USA 99:2061–2066
Labandeira CC, Sepkoski JJ Jr (1993) Insect diversity in the fossil record. Science 261:310–315
Labandeira CC, Wilf P, Johnson KR, Marsh F (2007b) Guide to insect (and other) damage types on compressed plant fossils. Version 3.0—Spring 2007). Smithsonian Institution, Washington. http://paleobilogy.si.edu/pdf:insectDamageGuide3.01.pdf
Lawrence JF, Ślipiński A (2013) Australian beetles: Morphology, classification and keys, vol 1. CSIRO, Collingwood
Lewis T (1973) Thrips: their biology, ecology and economic importance. Academic Press, London
López-Vaamonde C, Wikström N, Labandeira CC, Goodman S, Godfray HCJ, Cook JM (2006) Fossil-calibrated molecular phylogenies reveal that leaf-mining moths radiated millions of years after their host plants. J Evol Biol 19:1314–1326
Magallón S (2010) Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. Syst Biol 59:384–399
Marshall SA (2012) Flies: the natural history and diversity of Diptera. Firefly, Buffalo
McAlpine JF, Peterson BV, Shewell GE, Teskey HJ, Vockeroth JR, Wood DW (eds) (1981–1989) Manual of Nearctic Diptera vols 1–3. Canadian Government Publishing Centre, Hull, Quebec
McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. Proc Natl Acad Sci USA 106:7083–7088
McLoughlin S, Carpenter RJ, Jordan GJ, Hill RS (2008) Seed ferns survived the end-Cretaceous extinction in Tasmania. Am J Bot 95:465–471
McLoughlin S, Carpenter RJ, Pott C (2011) *Ptilophyllum muelleri* (Ettingsh.) comb. nov. from the Oligocene of Australia: last of the Bennettitales? Int J Plant Si 172:574–585
Miller NCE (1956) The biology of the Heteroptera. Leonard Hill, London

Moran NA, Tran P, Gerardo NM (2005) Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum *Bacteroidetes*. Appl Environ Microbiol 71:8802–8810

Naumann ID, Carne PB, Lawrence JF, Nielsen ES, Spradbery JP, Taylor RW, Whitten MJ, Littlejohn MJ (eds) (1991) The insects of Australia: A textbook for students and research workers, vols. 1, 2. Cornell, Ithaca

Nylin S, Wahlberg N (2008) Does plasticity drive speciation? Host-plant shifts and diversification in nymphaliine butterflies (Lepidoptera: Nymphalidae) during the Tertiary. Biol J Linn Soc 94:115–130

Paleobiology Database (2014) http://paleobiol.org. Last accessed 10 Feb 2014

Pellmyr O, Seagraves K (2003) Pollinator divergence within an obligate mutualism: two yucca moth species (Lepidoptera: Prodoxidae: *Tegeticula*) on the Joshua tree (*Yucca brevifolia*; Agavaceae). Ann Entomol Soc Am 96:716–722

Peñalver E, Labandeira CC, Barrón E, Delclòs X, Nel A, Nel P, Taffoureau P, Soriano C (2012) Thrips pollination of Mesozoic gymnosperms. Proc Natl Acad Sci USA 109:8623–8628

Rasnitsyn AP (1980) The origin and evolution of hymenopterous insects. Trans Paleontol Inst 174:1–192 (in Russian)

Rasnitsyn AP (1988) Principles and methods of phylogenetic reconstruction. In: Ponomarenko AG (ed) The Mesozoic-Cenozoic crisis in the evolution of insects. Nauka, Moscow, pp 191–207 (in Russian)

Rasnitsyn AP, Krassilov VA (2000) The first documented occurrence of phyllophagy in pre-Cretaceous insects: leaf tissues in the gut of Upper Jurassic insects from southern Kazakhstan. Paleontol J 34:301–309

Rasnitsyn AP, Quicke DLJ (eds) (2002) History of insects. Kluwer, Dordrecht

Ratzel SR, Rothwell GW, Mapes G, Mapes RH, Doguzhaeva LA (2001) *Pityostrobus hokodzensis*, a new species of pinaceous cone from the Cretaceous of Russia. J Paleontol 75:895–900

Ren D (1998) Flower-associated Brachycera flies as fossil evidence for Jurassic angiosperm origins. Science 280:85–88

Ren D (ed) (2010) Current research on palaeoentomology. Acta Geol Sin 84(4):655–1010

Ren D, Labandeira CC, Santiago-Blay JA, Rasnitsyn AP, Shih CK, Bashkuev A, Logan MAV, Hotton CL, Dilcher DL (2009) A probable pollination mode before angiosperms: Eurasian, long-proboscid scorpionflies. Science 326:840–847

Ross AJ, Jarzembowski EA (1993) Arthropoda (Hexapoda; Insecta). In: Benton MJ (ed) The fossil record 2. Chapman & Hall, London, pp 363–426

Schachat S, Labandeira CC, Gordon J, Chaney D, Levi S, Halthore MS, Alvarez J (2014) Plant–insect interactions from the Early Permian (Kungurian) Colwell Creek Pond, North-Central Texas: the early spread of herbivory in riparian environments. Int J Pl Sci 175: in press

Schuh RT, Slater JA (1995) True bugs of the world (Hemiptera: Heteroptera). Cornell, Ithaca

Sinitshenkova ND (2002) Ecological history of the aquatic insects. In: Rasnitsyn AP, Quicke DLJ (eds) History of insects. Kluwer, Dordrecht, pp 388–426

Sohn J-C, Labandeira CC, Davis D, Mitter C (2012) An annotated catalog of fossil and subfossil Lepidoptera (Insecta: Holometabola) of the world. Zootaxa 3286:1–132

Sukatcheva ID (1991) The Late Cretaceous stage in the history of the caddisflies (Trichoptera). Acta Hydroentom Lat 1:68–85

Taylor TN, Taylor EL, Krings M (2009) Paleobotany: the biology and evolution of fossil plants, 2nd edn. Elsevier, Amsterdam

Wang B, Szwedo J, Zhang H (2012a) New Jurassic Cercopoidea from China and their evolutionary significance (Insecta: Hemiptera). Palaeontology 55:1223–1243

Wang B, Zhang H, Jarzembowski EA (2013) Early Cretaceous angiosperms and beetle evolution. Front Pl Sci 4:360

Wang Y, Labandeira CC, Ding Q, Shih CK, Zhao Y, Ren D (2012b) An extraordinary Jurassic mimicry between a hangingfly and ginkgo from China. Proc Natl Acad Sci USA 109:20514–20519

Wang Y, Liu Z, Wang X, Shih C, Zhao Y, Engel MS, Ren D (2010) Ancient pinnate leaf mimesis among lacewings. Proc Natl Acad Sci USA 107:16212–16215
Wappler T, Labandeira CC, Rust J, Frankenhäuser H, Wilde V (2012) Testing for the effects and consequences of mid-Paleogene climate change on insect herbivory. PLoS ONE 7:e40744
Watson J (1977) Some Lower Cretaceous conifers of the Cheirolepidiaceae from the U.S.A. and England. Palaeontology 20:715–749
Weingartner E, Wahlberg N, Nylin S (2006) Dynamics of host plant use and species diversity in *Polygonia* butterflies (Nymphalidae). J Evol Biol 19:483–491
Wilf P, Labandeira CC, Johnson KR, Cúneo NR (2005) Richness of plant–insect associations in Eocene Patagonia: a legacy for South American biodiversity. Proc Natl Acad Sci USA 102:8944–8948
Wilf P, Labandeira CC, Johnson KR, Ellis B (2006) Decoupled plant and insect diversity after the end-Cretaceous extinction. Science 313:1112–1115
Winkler IS, Labandeira CC, Wappler T, Wilf P (2010) Diptera (Agromyzidae) leaf mines from the Paleogene of North America and Germany: implications for host use evolution and an early origin for the Agromyzidae. J Paleontol 84:935–954
Yeates DK, Wiegmann BM (eds) (2005) The evolutionary biology of flies. Columbia, New York
Zhang W, Shih CK, Labandeira CC, Sohn JC, Davis DR, Santiago-Blay JA, Flint O, Ren D (2013) New fossil Lepidoptera (Insecta: Amphiesmenoptera) from the Middle Jurassic Jiulongshan Formation of Northeastern China. PLoS ONE 8(11):e79500
Zherikhin VV, Mostovski MB, Vršanský P, Blagoderov VA, Lukashevich ED (1999) The unique lower Cretaceous locality Baissa and other contemporaneous fossil insect sites in North and West Transbaikalia. In: Proceedings of 1st International Palaeoentom Conference (Moscow, 1998). AMBA Projects, Bratislava, p 185–191
Zhou Z, Zhang B (1989) A sideritic *Protocupressinoxylon* with insect borings and frass from the Middle Jurassic, Henan, China. Rev Palaeobot Palynol 59:133–143

Chapter 14
The Evolution and Pollination of Oceanic Bellflowers (Campanulaceae)

Marisa Alarcón, Juan José Aldasoro, Cristina Roquet and Jens M. Olesen

Abstract Oceanic islands provide a good model for the study of species dispersal and evolution. We focus here on the evolution of pollination modes of oceanic island bellflowers (Campanulaceae), examining the degree of parallel evolution in different lineages of this family. Plants colonizing islands might either have experienced selective pressures on floral traits from vertebrate pollinators such as birds and lizards or have been pre-adapted to pollination by vertebrates prior to their colonization. The reconstruction of the ancestral pollination biology of Campanulaceae suggests that pollinators of the ancestors of bird-/lizard-pollinated bellflowers were insects. Moreover, in four island Campanulaceae lineages, only one was pre-adapted on the continent, and three made de novo shifts on the islands. Evolution towards bird pollination from insect-pollinated ancestors is also common in other island-groups, possibly because opportunistic birds are more efficient than insects. We review to what extent related species converge in their pollination ecology in related habitats on oceanic islands.

Marisa Alarcón and Juan José Aldasoro contributed equally to the study

M. Alarcón (✉) · J. J. Aldasoro
Institut Botànic de Barcelona (IBB-CSIC-ICUB), Passeig del Migdia s.n., Parc de Montjuïc, 08038 Barcelona, Spain
e-mail: malarcon@ibb.csic.es

C. Roquet
Laboratoire d'Écologie Alpine, CNRS UMR 5553, Université Grenoble, 1, BP 53, 38041 Grenoble Cedex 9, France

J. M. Olesen
Institute of Bioscience, Aarhus University, Ny Munkegade 114, 8000 Aarhus, Denmark

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_14,

14.1 Introduction

Oceanic islands are considered as 'natural laboratories' because they are relatively simple and isolated systems, whose geological age is usually known. They provide excellent opportunities for studying how species colonizes new areas, how they diversify, and how new biotic interactions influence their evolution. Animals and plants that disperse and establish on an oceanic island usually meet a community of native species that is different from the source continent; consequently, they may experience a release of antagonists or suffer constraints on new interactions.

Regarding plant–pollinator interactions, oceanic islands presents some unique properties: a scarcity of large flower-visiting insects (Anderson et al. 2001; Olesen and Jordano 2002; Olesen 2003; Olesen and Valido 2003a, b, 2004), a relatively high representation of small insects, mostly considered as generalist pollinators, and several vertebrate flower-visitors, such as bats, birds and lizards, which are also probably generalists (Olesen 1985; Olesen et al. 2002a, b; Dupont et al. 2003; Olesen 2003; Olesen and Valido 2004; Valido et al. 2004). Moreover, in oceanic islands, ornithophily and saurophily (i.e. pollination mediated by birds and lizards, respectively) have been documented for several plant species (Vogel et al. 1984; Olesen 1985; Olesen and Valido 2003a, b; Rodríguez-Rodríguez and Valido 2008; Ollerton et al. 2009).

According to the pollination syndrome concept, flowers of different plant species present similar characters, attracting flower-visiting animals of similar behaviour, morphology and physiology. Each syndrome is defined by a set of morphological and chemical attributes (e.g. flower robustness, orientation, odour, colour, shape and size, exposure of sexual organs and nectar composition and quantity), that partly reflect convergence caused by selective pressures exerted by similar animals, but that can be also limited by genetic/phylogenetic constraints (Faegri and Pijl 1979; Fenster et al. 2004). Here we focus on the origin and evolution of floral traits in oceanic bellflowers (Campanulaceae), a well-known group (Roquet et al. 2008, 2009; Mansion et al. 2012). Insect-pollinated bellflowers are generally white, yellow, blue or violet, and present small amounts of concentrated nectar rich in sucrose. Among them, bee-pollinated flowers may have a campanulate corolla, bee-fly- and butterfly-pollinated flowers are tubular, while flowers of selfers, generally, are small or inconspicuous. In contrast, bird-pollinated flowers, generally, are red-orange-yellow in colouration, scentless, present plenty of dilute nectar and have a robust texture (e.g. Rodríguez-Gironés 2004; Cronk and Ojeda 2008; Dalsgaard et al. 2008; Rodríguez-Rodríguez and Valido 2008).

Several hypotheses have been proposed to describe the origin and evolution of the saurophilous or ornithophilous flora of oceanic islands (Valido et al. 2004; Ojeda et al. 2012; Ortega-Olivencia et al. 2012), which may be summarized as follows:

(i) 'The island de novo specialist hypothesis' states that, first, mainland insect-pollinated plants established on the island, then, specialist nectar-feeding birds exerted selective pressures on flower traits promoting an evolution

towards ornithophily. Later, these birds became extinct and were substituted by pollinating opportunistic birds and lizards (or even insects).

(ii) 'The island de novo opportunist hypothesis' states that, after island colonization by mainland insect-pollinated ancestors, current opportunistic birds and lizards exerted selective pressures on some floral traits. Insects may also have been maintained as secondary visitors.

(iii) 'The relict hypothesis' affirms that the ancestors of island plants developed towards a specialized bird or lizard pollination syndrome in the mainland, before colonization of the islands. After colonization of islands, the specialists were replaced by generalist nectarivorous birds and lizards, which maintained a selection for some of the previously selected pressure traits. Insects may also become secondary visitors.

It is important to keep in mind that oceanic islands, being poor in species, are particularly vulnerable to biotic invasions and extinctions. Colonization events by birds and small insects are relatively frequent. Successful establishments have occurred many times during the island history, especially after their colonization by humans. Plasticity in plant–pollinator interactions and in pollination-associated characters may facilitate invasions of plants and animals into the native ecological networks (Olesen et al. 2002a; Kaiser-Bunbury et al. 2010; Danieli-Silva et al. 2012). For instance, a specialist island bird may have become extinct and substituted by other more opportunistic birds, such as passerines. Consequently, the selection pressure exerted by invaders may have changed in form and strength many times during the evolutionary history of island plants. Another fact, introducing a greater complexity into this scenario, is the existence of mixed insect–vertebrate pollination systems, which occur in islands that suffered temporal variation in pollinator fauna (Ortega-Olivencia et al. 2012).

14.2 The Bellflower Family

The bellflower family (Campanulaceae s. str.) is divided into three large tribes: Platycodoneae, Wahlenbergieae and Campanuleae (Fig. 14.1). The tribe Platycodoneae includes c. 90 species mainly distributed in Asia, and it is basal to the other two tribes (Roquet et al. 2008; Olesen et al. 2012; Wang et al. 2013), diverging from the rest of the family c. 36–40 mya (Olesen et al. 2012). The main exception (in terms of biogeographic distribution) within Platycodoneae is *Canarina*, which is found in East Africa and Macaronesia (one species is endemic to Canary Islands), thus showing a striking disjunction. Ancestors of *Canarina* probably dispersed from Asia to Africa during the Miocene, c. 13 mya (Mansion et al. 2012; Wang et al. 2013).

The separation between the ancestors of the two other tribes, Wahlenbergieae and Campanuleae, has been estimated to have occurred 25.9 mya. The Wahlenbergieae is essentially a southern hemisphere group that comprises 15 genera (Cupido et al. 2013).

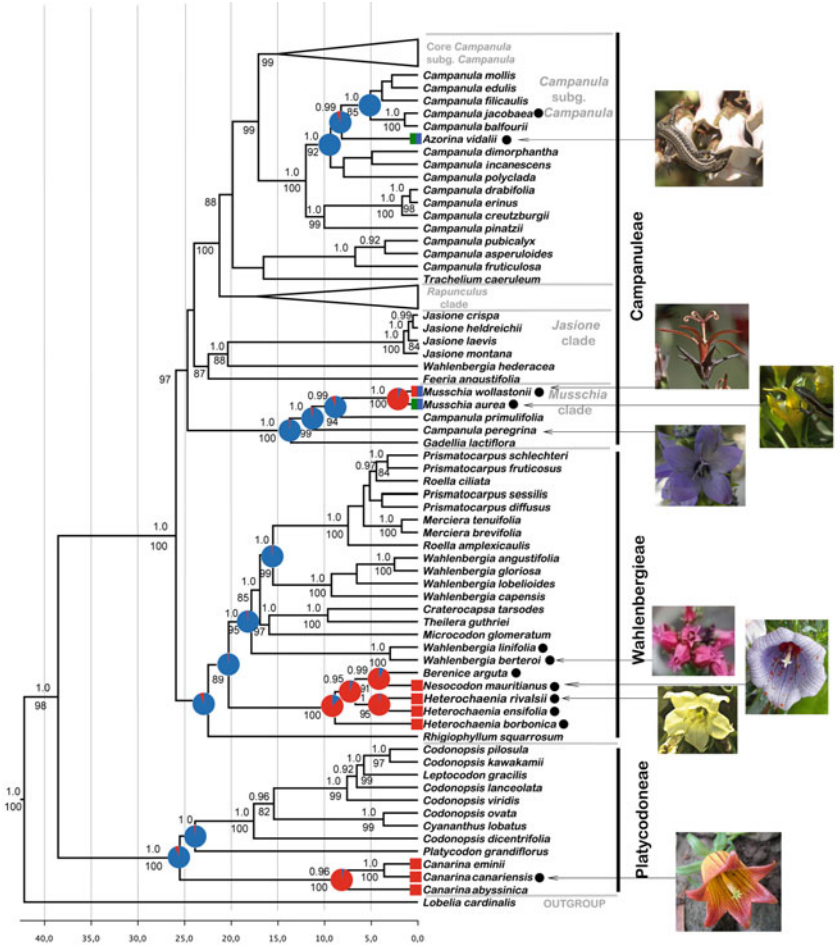


Fig. 14.1 Phylogenetic tree of Campanulaceae using five markers: *atp*B, *mat*K, *rbc*L, *pet*B-*pet*D and *trn*L-*trn*F. Trait state reconstruction of the pollination type of the bellflowers was carried out using maximum likelihood (Mesquite). The proportional likelihoods of the different character states in the ancestral reconstructions are indicated by the area red/blue in each pie diagram (red for bird/lizard pollination and blue for insect pollination). If a pie is entirely blue for a group, younger pies are not shown in order to simplify the figure. A red square at the branch tip indicates bird pollination, green and blue indicate lizards plus insects and red and blue indicate birds plus insects. Presence on oceanic islands is given by a black spot after the species name. The bootstrap values are above the branches in bold and posterior probabilities below (reanalysed from Olesen et al. 2012, with some new data)

Wahlenbergia, the largest and most widely distributed of the Wahlenbergieae genera, is monophyletic, including more than 170 species (Cupido and Conrad 1999; Welman and Cupido 2003; Prebble et al. 2011; Olesen et al. 2012; Cupido et al. 2013). Within Wahlenbergieae there are several oceanic endemics: several

Wahlenbergia species, *Berenice arguta*, *Heterochaenia rivalsii, H. borbonica,* and *H. ensifolia*, and *Nesocodon mauritianus*. Another species, *Heterochaenia fragrans*, was described from Réunion recently (Thomas et al. 2008). The *Heterochaenia–Berenice–Nesocodon* group split from the rest of the species of Wahlenbergieae 20.3 mya.

The largest group of Campanulaceae is Campanuleae, with more than 600 species which mostly grow in the Old World. They can be divided into four nested clades: (i) the *Campanula lactiflora-Musschia* clade, (ii) the *Jasione-Feeria* clade, (iii) the *Rapunculus* clade and the iv) *Campanula* s. str. clade. The ancestor of the clade of *Musschia* and the *Campanula peregrina* group appeared 13.4 mya, possibly in the Mediterranean basin (Roquet et al. 2009), while the ancestor of *Musschia* appeared 8.8 mya. The large *Campanula* s. str. clade includes *Azorina, Diosphaera, Edrajanthus, Michauxia* and many species of *Campanula.* It includes a small subclade formed by 20 African and 3 Macaronesian species. This group diversified from the late Miocene onwards, three dispersal events brought the ancestor of *Azorina* to the Azores 8.3 mya, that of *C. occidentalis* c. 5 mya to Canary Islands and that of *Campanula jacobaea* to the Cape Verde Islands 1.2 mya (Alarcón et al. 2013).

14.3 Comparison of the Lineages of Oceanic Bellflowers

Endemic oceanic bellflowers are found in tropical and subtropical islands that vary substantially in geological age (Table 14.1). Thus, the ancestors of oceanic endemic bellflowers probably arrived at different ages and following different dispersal routes: the Mascarenes are an archipelago of three large islands, Mauritius, Réunion and Rodrigues, and several small islets surrounding Mauritius, where it has been documented that many animal species arrived either from Africa and Madagascar or with the Equatorial current and jumping along a string of now submerged islands from SE Asia and Australia. The Juan Fernández Archipelago is formed by two major islands: Masatierra and Masafuera, and its main colonization source were most likely the coasts of Chile. In the case of Macaronesian archipelagos, there are several seamounts closer to Africa, likely former islands, which could have constituted stepping stones for many taxa. We present below seven bellflower genera including endemic taxa to these islands, which belong to four clearly distinct lineages showing a shift in pollination mode. We compare these taxa to their closest mainland relatives in terms of pollination and other reproductive aspects, integrating this information in an evolutionary and biogeographic context.

14.3.1 Canarina

The first case of ornitophily in the evolution of Campanulaceae occurred in the genus *Canarina,* which belongs to the tribe Platycodoneae, the basal group of

Table 14.1 Island age for the studied archipelagos

Archipelago	Island age	References
Canary Is.	Fuerteventura 21 my, Lanzarote c. 16 my, Gran Canaria c. 15 my Tenerife c. 7.5 my, La Gomera 14 my, La Palma 3 my, El Hierro 1.5 my	Anguita et al. (2002), Carracedo et al. (2002)
Mascarenes Is.	Mauritius 7.8 my, Réunion 2.1 my, Rodrigues 1.5–15.0 my	Fisk et al. (1989), Duncan and Storey (1992), Warren et al. (2003)
Juan Fernandez Is.	Masafuera 1–2.4 my, Masatierra 3.8–4.2 my, Santa Clara 5.8 my	Stuessy et al. (1984)
Madeira Is.	Madeira c. 4.6 my, Porto Santo c. 14 my, Desertas c. 3.6 my	Feraud et al. (1981), Mitchell-Thomé (1985), Ferreira et al. (1988)
Cape Verde Is	Sal and Boavista 26 my, Maio 7–20 my, Brava 5.5 my, S. Nicolau 6.5 my, S Antao c.7 my, Santiago 10 my	Mitchell-Thomé (1985), Plesner et al. (2003), Duprat et al. (2007)
Azores Is.	Santa Maria 8.1 my and Sao Miguel 4.0 my, Terceira 3.5 my, Graciosa 2.5 my, Sao Jorge 0.6 my, Pico 0.3 my, Faial 0.7 my, Flores 2.2 my, Corvo 0.7 my	Abdel-Monem et al. (1975), Borges and Hortal (2009)

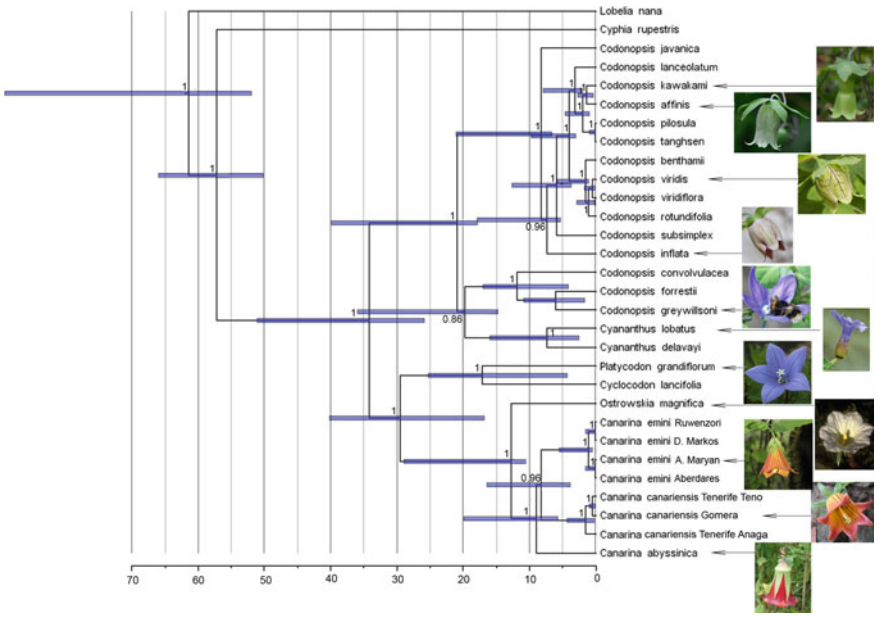


Fig. 14.2 Dated phylogenetic tree obtained using Beast and based on *cp*DNA sequences (*trn*L–*trn*F, *trn*S–*trn*G and *pet*B–*pet*D) of the Platycodoneae clade. An upper age constraint of 59 mya was deduced from the data of Bell et al. (2010), who calculated the age of many clades of the Angiosperm tree by calibrating it with fossil constraints. The ages obtained by these authors ranged between 64 and 41 mya for the separation of Campanulaceae and Lobeliaceae. Numbers beside the nodes indicate posterior probabilities

Campanulaceae (Fig. 14.2). The ancestor of *Canarina* dispersed from Asia to Africa and likely evolved into several species that colonized eastern Africa and later the Canary Islands, c. 8 mya (Roquet et al. 2009; Mansion et al. 2012; Olesen et al. 2012). This genus is constituted by three species: the East African *Canarina eminii,* a drought tolerant epiphyte which requires high light intensity and is found on the outer portions of the canopy (twig epiphyte) (Alemayehu 2006). In contrast, the two congeners: *Canarina canariensis,* which grows in the Canary Islands, and *Canarina abyssinica*, distributed in East Africa, are terrestrials. The latter species has tendrils (Hedberg 1961). Both, *Canarina canariensis* and *Canarina abyssinica*, may climb on neighbouring plants.

Other Platycodoneae such as *Codonopsis clematidea* and *C. pilosula* are pollinated by wasps and present more diluted nectar than the rest of *Codonopsis*. However, the species in the *C. grey-willsonii* group are pollinated by bumblebees. *Cyananthus* is pollinated by generalist insects like flies, small bees, etc., while *Platycodon* and *Cyclocodon* are pollinated by bumblebees. The large flowers of the monospecific genus *Ostrowskia* suggest that they may attract bats or moths as pollinators, but this species has still not been studied in the field. The three *Canarina* species are pollinated by birds and lizards (Vogel et al. 1984; Rodríguez-Rodríguez

and Valido 2011; Olesen et al. 2012). In Africa *C. eminii* and *C. abyssinica* are pollinated by Nectariniidae birds, while in the Canary Islands *C. canariensis* is pollinated by generalist birds, e.g. members of the Sylviidae and Paridae (Table 14.2; Olesen 1985; Rodríguez-Rodríguez and Valido 2011; Olesen et al. 2012; Padilla et al. 2012). The berries of *Canarina* have a sticky sweet pulp, and in the Canary Islands those of *C. canariensis* are consumed by birds and lizards (Olesen 1985; Olesen et al. 2012). Thus these animals become double mutualists, serving *C. canariensis* as both pollinators and seed dispersers. Monkeys have been observed eating the plants of *C. eminii* (Yamagiwa et al. 2005; Yamagiwa and Basabose 2006; Davenport et al. 2010). In contrast, the ripe fruits of *Ostrowskia* are dry capsules and the seeds have a small wing. These data suggest that the migration to Africa was followed by a shift to pollination and seed dispersal by birds. A few other Platycodoneae (e.g. species of *Cyclocodon* and *Campanumoea*) have fleshy fruits that are likely consumed by birds.

14.3.2 Heterochaenia, Nesocodon and Wahlenbergia

In the tribe of Wahlenbergieae there are three genera with endemic oceanic species: *Heterochaenia*, *Nesocodon* and *Wahlenbergia*. *Heterochaenia* and *Nesocodon* constitute a clade together with *Berenice* (Fig. 14.1). The *Heterochaenia–Berenice–Nesocodon* group split from *Wahlenbergia* and the rest of the species of Wahlenbergieae *s. str.* 20.3 mya (Olesen et al. 2012). The ancestor of *Nesocodon* and *Berenice* diverged from *Heterochaenia rivalsii* and *H. ensifolia* c. 6.7 mya. The Mascarene *Heterochaenia* and *Nesocodon* species may have originated in the wet East African or Malagasy forests. Mascarene bellflowers are found in the mountains, from 400 m (*N. mauritianus* in Mauritius, which has red nectar) to 1700–2400 m (*Heterochaenia* in Réunion Is.). These four species are visited by vertebrates, although the genus *Berenice*, sister to *Nesocodon*, is autogamous or pollinated by insects. Two bird species have been visiting the flowers of *N. mauritianus*: the introduced Red-whiskered Bulbul *Pycnonotus jocosus* (L.) and the Mauritian Merle *Hypsipetes olivaceus* (Jardine & Selby) (Table 14.2). The flowers of *Heterochaenia rivalsii* and *H. ensifolia* flowers are visited by the Mascarene Grey White-eye *Zosterops borbonicus* (Olesen et al. 2012).

Oceanic *Wahlenbergia* species, such as those growing in New Zealand (Lord 2008) or on Juan Fernandez archipelago, are autogamous or pollinated by insects. *Wahlenbergia berteroi* from the Robinson Crusoe Island shows floral traits that may suggest bird pollination, but only flies and ants have been observed as visitors (Anderson et al. 2000, 2001; Bernardello et al. 2000, 2001). Most of the studied continental species of *Wahlenbergia* and related genera are also pollinated by insects (Gess 1999; Peter and Johnson 2008; Campbell et al. 2012; Welsford and Johnson 2012; Uys and Cron 2013).

Table 14.2 Some island interactions and the origin of pollination traits

Families and species	Islands	Visitors and pollinators (◆ introduced or invasive)			Type of adaptation to pollinators: relict or de novo (*mixed vertebrate-insect system)	Reference
		Birds	Lizards	Insects		
ARALIACEAE						
Gastonia mauritiana	Mauritius Is.		*Phelsuma ornata*		relict or de novo	Nyhagen et al. (2001)
BORAGINACEAE						
Echium wildpretii	Canary Is.	*Serinus canarius*, *Phylloscopus canariensis*,	–	*Eucera gracilipes* *Anthophora alluaudi* *Apis mellifera* ◆ *Hylaeus canariensis* *Lasioglossum chalcodes* *L. viride* *Megachile canariensis* *Anthophora alluaudii* *Melecta curvispina* *Macroglossum stellatarum* *Anastoechus latifrons* *Anthrax anthrax*	de novo (*)	Valido et al. (2002)
CAMPANULACEAE						
Canarina canariensis	Canary Is.	*Sylvia conspicillata*, *S. melanocephala*, *Cyanistes teneriffae*, *Phylloscopus canariensis*,	–	*Lasioglossum viride*	Relict (sister species in E Africa)	Rodríguez-Rodríguez and Valido (2011), Olesen et al. (2012)

(continued)

Table 14.2 (continued)

Families and species	Islands	Visitors and pollinators (◆ introduced or invasive)			Type of adaptation to pollinators: relict or de novo (*mixed vertebrate-insect system)	Reference
		Birds	Lizards	Insects		
Musschia wollastonii	Madeira Is.	*Sylvia atricapilla*	–	*Bombus maderensis* *Artogeia rapae* *Lycaena phlaeas*	de novo (*)	Olesen and Valido (2003b)
Musschia aurea	Madeira Is.	–	*Lacerta dugesii*	*Bombus maderensis*	de novo (*)	Olesen and Valido (2003a)
Azorina vidalii	Azores Is.	–	*Lacerta dugesii* ◆ (in Terceira and S[a] Maria)	*Apis mellifera* ◆ (all islands) *Macroglossum stellatarum*	de novo (*)	Olesen et al. (2012)
Heterochaenia borbonica	Réunion Is.	*Zosterops*?	–	–	de novo	Olesen et al. (2012)
Heterochaenia ensifolia	Réunion Is.	*Zosterops borbonicus*	–	–	de novo	Olesen et al. (2012)
Heterochaenia rivalsii	Réunion Is.	*Zosterops borbonicus*	–	–	de novo	Olesen et al. (2012)
Nesocodon mauritianus	Mauritius Is	*Hypsipetes olivaceus* *Pycnonotus jocosus*◆	–	–	de novo	Olesen et al. (2012)
CRASSULACEAE						
Aeonium arboreum	Canary Is.	*Serinus canarius* *Phylloscopus collybita*	–	–	de novo	Valido et al. (2004)

(continued)

Table 14.2 (continued)

Families and species	Islands	Visitors and pollinators (◆ introduced or invasive)			Type of adaptation to pollinators: relict or de novo (*mixed vertebrate-insect system)	Reference
		Birds	Lizards	Insects		
FABACEAE						
Lotus maculatus	Canary Is.	*Phylloscopus canariensis* *Cyanistes caeruleus*	–	–	de novo(*)	Ojeda et al. (2012, 2013)
Lotus berthelotii	Canary Is.	*Sylvia melanocephala* *Serinus canarius* *Cyanistes teneriffae,*	–	–	de novo	Ojeda et al. (2012, 2013)
Lotus pyranthus	Canary Is.	*Sylvia melanocephala* *Serinus canarius* *Cyanistes teneriffae,*	–	–	de novo	Ojeda et al. (2012, 2013)
Sophora fernandeziana	Juan Fernández Is.	*Sephanoides fernandensis,* *S. sephaniodes*	–	–	relict	Bernardello et al. (2001)
LAMIACEAE						
Cuminia eriantha	Juan Fernández Is.	*Sephanoides fernandensis,* *S. sephaniodes*	–	–	relict	Bernardello et al. (2001)
LILIACEAE						

(continued)

Table 14.2 (continued)

Families and species	Islands	Visitors and pollinators (◆ introduced or invasive)			Type of adaptation to pollinators: relict or de novo (*mixed vertebrate-insect system)	Reference
		Birds	Lizards	Insects		
Aloe mayottensis	Mayotte Is.	*Nectarinia coquereli*	–	–	relict	Pailler et al. (2002)
Lomatophyllum tormentorii	Mauritius Is.	–	*Phelsuma ornata*	–	relict	Nyhagen et al. (2001)
MALVACEAE						
Navaea phoenicea	Canary Is	*Sylvia melanocephala*,	–	*Bombus canariensis* *Pararge siphioides* *Ancistrocerus haematodes*	relict	González and Fuertes (2011), Fuertes-Aguilar et al. (2002)
		S. atricapilla *Phylloscopus canariensis*, *Cyanistes teneriffae*		*Apis mellifera* ◆		
Trochetia granulata	Réunion Is.	*Zosterops borbonicus* *Z. olivaceus*	*Phelsuma borbonica*	*Apis mellifera* ◆	de novo	Le Péchon et al. (2010, 2013)
Trochetia blackburniana	Mauritius Is.	*Zosterops chloronothos*	*Phelsuma cepediana*	–	de novo	Hansen et al. (2007)
ORCHIDACEAE						
Angraecum striatum	Réunion Is.	*Zosterops borbonicus*	–	–	de novo	Micheneau et al. (2006)
PLANTAGINACEAE						

(continued)

Table 14.2 (continued)

Families and species	Islands	Visitors and pollinators (◆ introduced or invasive)			Type of adaptation to pollinators: relict or de novo (*mixed vertebrate-insect system)	Reference
		Birds	Lizards	Insects		
Isoplexis canariensis	Canary Is	*Sylvia melanocephala*, *Serinus canarius*, *Phylloscopus canariensis*, *Cyanistes caeruleus*, *Fringilla coelebs*	*Gallotia galloti*	–	relict or de novo	Dupont et al. (2004), Rodríguez-Rodríguez et al. (2013)
Isoplexis isabelliana	Canary Is.	*Phylloscopus canariensis*	–	–	relict or de novo	Valido et al. (2004)
Isoplexis chalcantha	Canary Is.	*Phylloscopus canariensis*	–	–	relict or de novo	Valido et al. (2004)
Isoplexis sceptrum	Madeira Is	*Sylvia atricapilla*	–	–	relict or de novo	Olesen and Valido (2003b)
Scrophularia calliantha	Canary Is.	*Sylvia melanocephala*, *S. atricapilla* *Cyanistes teneriffae* *Phylloscopus canariensis*	*Gallotia stehlini*	*Lasioglossum viride*	relict(*)	Ortega-Olivencia et al. (2012), Navarro-Pérez et al. (2013)
ROUSSEACEAE						
Roussea simplex	Mauritius Is.	*Zosterops mauritianus*	*Phelsuma cepediana*	–	de novo	Hansen and Muller (2009)

14.3.3 Musschia

Musschia is a genus formed by 2–3 species (*M. aurea* and *M. wollastonii*) endemic to the Madeiran archipelago; a third species was described from the Desertas recently (Menezes de Sequeira et al. 2007). The ancestor of *Musschia* and its sister clade (the *Campanula peregrina* group) appeared 6.8 mya, possibly in the W Mediterranean–N Africa area (Roquet et al. 2009), while the ancestor of *Musschia* probably appeared in Madeira 4.1 mya. The two main *Musschia* species would have split 1.5 mya. While the *Campanula peregrina* group presents rotate, whitish or bluish flowers, which attract generalist insects, *M. wollastonii* and *M. aurea* are pollinated by vertebrates, 'viz. birds and lizards, respectively. Concretely, *Musschia wollastonii* is visited by *Sylvia atricapilla*, but also by many insects (e.g. Apidae: *Bombus maderensis*, Pieridae: *Artogeia rapae* L., Lycaenidae: *Lycaena phlaeas* L.,Vespidae, Syrphidae and Muscidae) (Table 14.2). However, most of these insects are only nectar thieves and not pollinators. *M. aurea* is visited and pollinated by *Lacerta dugesii* and *Bombus maderensis. Lacerta dugesii* colonized the Madeiran Archipelago 2.8 mya (Brehm et al. 2003). The first colonization of *Sylvia atricapilla* might have occurred 3 mya (Dietzen et al. 2008). This is a late colonization compared to the age of the islands. Given that the colonization of *Musschia* is much older than their current pollinators, the *Musschia* ancestor must have relied on other pollinators, maybe migrant birds or (nowadays extinct) lizards.

14.3.4 Azorina and Campanula jacobaea

Basal to the main Afro-Macaronesian clade of *Campanula* species is *Azorina vidalii*, which is isolated phylogenetically and biogeographically. *Azorina vidalii* (Azores Is.) and a group of c. 22 Afro-Macaronesian species are embedded in a subclade within the *Campanula s. str.* clade. The species is thought to be derived from an ancient lineage related to the Asian *C. dimorphantha* and its relatives (Mansion et al. 2012). The ancestor of *A. vidalii* and the rest of the Afro-Macaronesian bellflowers diverged from its Central-Asian sister clade (Fig. 14.1; Schaefer et al. 2011; Mansion et al. 2012; Alarcón et al. 2013). At this time, other extinct *Azorina* relatives could have grown in different parts of northern Africa, while the Azores Is. served as a refuge for *Azorina.* Thus, two different dispersal events might have brought the ancestors of *Azorina* to the Azores 4.2 mya and *C. jacobaea* to the Cape Verde Islands 1.4 mya. Another subclade of species, the mostly autogamous *C. kremeri* and *C. occidentalis* are distributed in Morocco and the Iberian Peninsula and the Canary Islands, respectively, which stresses the high dispersal capability of bellflowers. *Azorina vidalii* is a small shrub, endemic to the Azores archipelago. Here, it grows on eight of the nine islands. Its habitat is rocky sea cliffs and deposits. All populations are visited by insects, but in Santa Maria

and Terceira, the lizard *Lacerta dugesii* was also observed. This species mediates cross-pollination. Several insect species visit the plant in the rest of the archipelago, but only the honeybee *Apis mellifera* was seen carrying and depositing pollen (Table 14.2). *Campanula jacobaea* from the Cape Verde Islands was always visited by small solitary Halictidae bees, but in some populations, the visitation rate is very low and self-pollination occurs (Alarcón et al. 2013). Finally, *C. occidentalis* is autogamous but also pollinated by small insects.

14.4 Traits Associated with Bird/Lizard Pollination

In Campanulaceae, flowers of lizard-pollinated species are often smaller than those of bird-pollinated ones, except for flowers of *Heterochaenia*. Flower orientation and shape are pendant and bell-shaped, respectively, except for *Musschia,* which has an upright and in the upper part rotated flower. Lizard-pollinated flowers have scent, whereas bird-pollinated ones are scentless. Nectar volume varies from 9 to 92 μl and the concentration is weak, 8–22 %, whereas the nectar of insect-pollinated species usually is higher than 30 %. However, there is a wide variation in nectar concentration among bird- and lizard-visited species. Also, there is a wide variation in flower size among bird- and lizard-visited species and nectar concentration and flower size are negatively correlated (Nectar conc. $= -32.7 \log_{10}$(Corolla tube length) $+ 66.9$, $R^2_{\text{adj.}} = 0.35$, $F = 5.82$, $P < 0.04$; Fig. 14.3; Olesen et al. 2012). The site of secondary pollen presentation is the style in *Azorina*, *Nesocodon* and *Canarina,* on both style and part of the stigma in *Heterochaenia*, and on the stigma in *Musschia*. Dichogamy is weak in *Musschia* and *Heterochaenia*, and strong in *Azorina, Nesocodon* and *Canarina.* Usually, plants pollinated by birds or lizards present larger flowers and more diluted nectar than those pollinated by insects, but there is a wide range in robustness, size, colours, shape and nectar concentration across studied bellflowers. These trends suggest weak convergence in floral biology in most island bellflowers.

14.5 Plasticity and Conservatism in the Islands Plant–Visitor Interactions

Convergent evolution is an important evolutionary process both in mainland and on islands, as many unrelated taxa show similar adaptations to similar selective agents. ‘Pollination syndromes’, e.g. ornitophily, saurophily and melittophily, are examples of convergent evolution. Though pollination syndromes are the result of convergent evolution across different taxonomic groups, they do not really predict floral visitors, but only describe suites of traits that have evolved in a more or less correlated way under the selection from principal pollinators, but also influenced

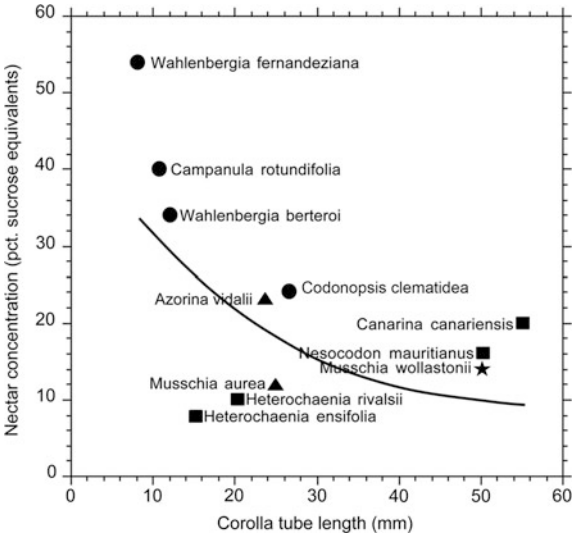


Fig. 14.3 Correlation between nectar concentration and corolla length for the studied Campanulaceae and two additional *Wahlenbergia* species. Symbols indicate pollinator types: spots = insects, squares = birds, triangles = lizards and insects, and stars = birds and insects. Data for *Campanula rotundifolia* L. are from Creswell and Robertson (1994) and those for *Wahlenbergia berteroi* and *W. fernandeziana* Skottsb. are from Bernardello et al. (2000) (after Olesen et al. 2012, including new data on *Codonopsis clematidea*)

by secondary floral visitors or nectar and pollen robbers (Thomson 2003; Martén-Rodríguez et al. 2010). Consequently, there is currently a debate about pollination syndromes, centred around the utility of the concept in predicting pollinator guilds, its accuracy in plant assignment to specific syndromes, and the role of pollinators in shaping these syndromes. Besides, rapid changes in island faunas, extinction of specialist pollinators, presence of generalized species (one plant species interacting with many animals and vice versa) and mixed insect–vertebrate pollination systems might blur the convergence of characters.

Birds that visit flowers with a 'mixed vertebrate-insect system' are attracted by easily accessible, non-restrictive flowers, with a type of nectar that fulfil their physiological requirements, and these flowers are also visited by other pollinator groups (lizards, bats, insects). This latter kind of 'syndrome' seems more common than previously thought in islands (Ojeda et al. 2012; Ortega-Olivencia et al. 2012). Fluctuations in the pollinator assemblages in oceanic islands results in a fluctuating selection favouring a generalized flower biology. Minor changes in corolla robustness, shape and size, for instance, may attract new groups of floral visitors. Thus, animals, after establishing on islands and initiating interactions with present plants may modify a few reproductive traits in order to establish a successful mutualism. Although plants have numerous different strategies to adapt to changing pollination regimes, not all plant populations are equally likely to undergo adaptive selection in response to new visitors and pollinators. These differences are patent in

the mentioned bellflower groups. The adaptive path taken may change among different groups, species or even among different populations of the same species, depending on the prevailing ecological, demographic, genetic and phylogenetic factors. Moreover, some characters presumably are part of an adaptive 'fail-safe' mechanism, that attracts secondary pollinators, when the primary pollinators are scarce or absent.

Most of our study islands have a relatively stable climate that may allow an accumulation of biodiversity (i.e. species and their links) and convert them into important refuges of diversity. However, changes in atmospheric circulation, sea-level changes and anthropogenic disturbance mediated many invasions, extinctions and changes in the flora and fauna of oceanic islands at the end of the tertiary and quaternary (Cronk 1997). As a consequence, the relationships between vertebrates and plants might have changed many times during this period. This may be the evolutionary background for the great lability in plant–bird and plant–lizard relationships, observed among the set of island bellflowers.

14.6 Conclusions

Plants that colonize islands might either have experienced a flower-selection pressure from birds and lizards, or have been pre-adapted to pollinating vertebrates prior to their colonization, or both. The reconstruction of the ancestral pollination biology of Campanulaceae suggests that pollinators of the ancestors of bird-/lizard-pollinated island bellflowers were insects, and in general bees (Olesen et al. 2012). In the case of continental Campanulaceae, the extant bird pollination syndrome may be a 'relict', because during the glacial and dry periods of the Neogene, some mainland species may have disappeared from large parts of their range, and have only survived in mesic refuges such as oceanic islands, or they may have adapted 'de novo' to the vertebrate visitation (Valido et al. 2004). Vertebrate pollination evolved independently at least four times in Campanulaceae, and these shifts seem in three of the cases associated with island colonization. However, a discrimination between the two first hypotheses on the origin of each pollination syndrome (e.g. the island de novo specialist hypothesis, and the island de novo opportunistic hypothesis) is difficult, because we neither know the complete history of each relationship, because many sisters to current species are extinct nor do we know the degree of specialization of the relationship.

Whereas a relict condition was interpreted for taxa such as *Canarina, Lavatera*, *Aloe, Sophora, Cuminia*, etc. (Table 14.2), many studies are increasing the list of island ornithophilous taxa that are better explained by the 'de novo' hypotheses (e.g. *Musschia*, *Heterochaenia*, *Nesocodon*, *Lotus*, *Angraecum*, *Roussea*, etc.; Table 14.2; Ojeda et al. 2012; Olesen et al. 2012). Evolution towards bird pollination from insect-pollinated ancestors seems to be the norm in oceanic islands, possibly because opportunistic birds (when they are present) are usually more efficient pollinators than insects (Dupont et al. 2004).

References

Abdel-Monem AA, Fernandez LA, Boone GM (1975) K-Ar ages from the eastern Azores group (Santa Maria, São Miguel and the Formigas islands). Lithos 8:247–254

Alarcón M, Roquet C, García-Fernández A, Vargas P, Aldasoro JJ (2013) Phylogenetic and phylogeographic evidence for a Pleistocene disjunction between *Campanula jacobaea* (Cape Verde islands) and *C. balfourii* (Socotra). Mol Phylogenet Evol 69:828–836

Alemayehu T (2006) Diversity and ecology of vascular epiphytes in Harenna Afromontane forest, Bale, Ethiopia. Ph.D. Thesis, Addis Ababa University

Anderson GJ, Bernardello G, Lopez PS, Crawford DJ, Stuessy TF (2000) Reproductive biology of *Wahlenbergia* (Campanulaceae) endemic to Robinson Crusoe island (Chile). Plant Syst Evol 223:109–123

Anderson GJ, Bernardello G, Stuessy TF, Crawford DJ (2001) Breeding system and pollination of selected plants endemic to Juan Fernandez islands. Am J Bot 88:220–233

Anguita F, Márquez A, Castiñeiras P, Hernán F (2002) Los volcanes de Canarias. Guía geológica e itinerarios. Ed Rueda, Madrid

Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-visited. Am J Bot 97:1296–1303

Bernardello G, Galetto L, Anderson GJ (2000) Floral nectary structure and nectar chemical composition of some species from Robinson Crusoe island (Chile). Can J Bot 78:862–872

Bernardello G, Anderson GJ, Stuessy TF, Crawford DJ (2001) A survey of floral traits, breeding systems, floral visitors, and pollination systems of the angiosperms of the Juan Fernandez islands (Chile). Bot Rev 67:255–308

Borges PA, Hortal J (2009) Time, area and isolation: factors driving the diversification of Azorean arthropods. J Biogeog 36:178–191

Brehm A, Jesus J, Spinola H, Alves C, Vicente L, Harris DJ (2003) Phylogeography of the Madeiran endemic lizard *Lacerta dugesii* inferred from mtDNA sequences. Mol Phylogenet Evol 26:222–230

Campbell DR, Bischoff M, Lord JM, Robertson AW (2012) Where have all the blue flowers gone: pollinator responses and selection on flower colour in New Zealand *Wahlenbergia albomarginata*. J Evol Biol 25:352–364

Carracedo JC, Perez-Torrado FJ, Ancochea E, Meco J, Hernán F, Cubas CR, Casillas R, Rodríguez-Badiola E, Ahijado A (2002) Cenozoic volcanism II: the Canary islands. In: Gibbons W, Moreno T (eds) The Geology of Spain. Geolog Soc London, London, pp 439–472

Cresswell JE, Robertson AW (1994) Discrimination by pollen-collecting bumblebees among differentially rewarding flowers of an alpine wildflower *Campanula rotundifolia* L. (Campanulaceae). Oikos 69:304–308

Cronk Q (1997) Islands, stability, diversity, conservation. Biodiv Conserv 6:477–493

Cronk Q, Ojeda I (2008) Bird-pollinated flowers in an evolutionary and molecular context. J Exp Bot 59:715–727

Cupido CN, Conrad F (1999) Bellflowers-Getting to know the South African bellflowers. Veld Flora 85:180–181

Cupido CN, Prebble JM, Eddie WM (2013) Phylogeny of Southern African and Australasian Wahlenbergioids (Campanulaceae) based on ITS and *trn*L-F sequence data: implications for a reclassification. Syst Bot 38:523–535

Dalsgaard B, Martín AMG, Olesen JM, Timmermann A, Andersen LH, Ollerton J (2008) Pollination networks and functional specialization, a test using Lesser Antillean plant–hummingbird assemblages. Oikos 117:789–793

Danieli-Silva A, de Souza JMT, Donatti AJ, Campos RP, Vicente-Silva J, Freitas L, Varassin IG (2012) Do pollination syndromes cause modularity and predict interactions in a pollination network in tropical high-altitude grasslands? Oikos 121:35–43

Davenport TR, De Luca DW, Bracebridge CE, Machaga SJ, Mpunga NE, Kibure O, Abeid YS (2010) Diet and feeding patterns in the kipunji (*Rungwecebus kipunji*) in Tanzania's Southern Highlands: a first analysis. Primates 51:213–220

Dietzen C, García-del-Rey E, Castro GD, Wink M (2008) Phylogenetic differentiation of *Sylvia* species (Aves: Passeriformes) of the Atlantic islands (Macaronesia) based on mitochondrial DNA sequence data and morphometrics. Biol J Linn Soc 95:157–174

Duncan RA, Storey M (1992) The life cycle of Indian Ocean hotspots. Am Geophys Union Monogr 70:91–103

Dupont YL, Hansen DM, Olesen JM (2003) Structure of a plant–flower-visitor network in the high-altitude sub-alpine desert of Tenerife, Canary islands. Ecography 26:301–310

Dupont YL, Hansen DM, Rasmussen JT, Olesen JM (2004) Evolutionary changes in nectar sugar composition associated with switches between bird and insect pollination: the Canarian bird-flower element revisited. Funct Ecol 18:670–676

Duprat HI, Friis J, Holm PM, Grandvuinet T, Sørensen RV (2007) The volcanic and geochemical development of São Nicolau, Cape Verde islands: constraints from field and $^{40}Ar/^{39}Ar$ evidence. J Volcanol Geotherm Res 162:1–19

Faegri K, van der Pijl L (1979) The principles of pollination ecology, 2nd edn. Pergamon Press, Oxford

Fenster CB, Armbruster WS, Wilson P, Dudash MR, Thomson JD (2004) Pollination syndromes and floral specialization. Annu Rev Ecol Evol Syst 35:375–403

Feraud G, Gastaud J, Schmincke HU, Pritchard G, Lietz J, Bleil U (1981) New K-Ar ages, chemical analyses and magnetic data of rocks from the islands of Santa Maria (Azores), Porto Santo and Madeira (Madeira Archipelago) and Gran Canaria (Canary islands). Bull Volcanol 44:359–375

Ferreira MP, Macedo CR, Ferreira JF (1988) K-Ar geochronology in the Selvagens, Porto Santo and Madeira islands (Eastern Central Atlantic): A 30 my spectrum of submarine and subaerial volcanism. Abstr Lunar Planet Sci Conf 19:325–326

Fisk MR, Duncan RA, Baxter AN, Greenough JD, Hargraves RB, Tatsumi Y (1989) Réunion hotspot magma chemistry over the past 65 my: results from Leg 115 of the ocean drilling program. Geology 17:934–937

Fuertes-Aguilar J, Ray MF, Francisco-Ortega J, Santos-Guerra A, Jansen RK (2002) Molecular evidence from chloroplast and nuclear markers for multiple colonisations of *Lavatera* (Malvaceae) in the Canary islands. Syst Bot 27:74–83

Gess SK (1999) Wasps and bees, stars and bells: an intricate pollination study of *Wahlenbergia*. Veld and Flora 85:80–81

González Fernández de Castro A, Fuertes Aguilar J (2011) Ecología y evolución de plantas ornitófilas de la Macaronesia. El indiferente 21:64–75

Hansen DM, Müller CB (2009) Reproductive ecology of the endangered enigmatic Mauritian endemic *Roussea simplex* (Rousseaceae). Int J Plant Sci 170:42–52

Hansen DM, Kiesbüy HC, Jones CG, Müller CB (2007) Positive indirect interactions between neighboring plant species via a lizard pollinator. Am Nat 169:534–542

Hedberg O (1961) Monograph of the genus *Canarina* L. (Campanulaceae). Svensk Bot Tidskr 55:17–62

Kaiser-Bunbury CN, Traveset A, Hansen DM (2010) Conservation and restoration of plant–animal mutualisms on oceanic islands. Perspect Plant Ecol Evol Syst 12:131–143

Le Péchon T, Dubuisson JY, Haevermans T, Cruaud C, Couloux A, Gigord LD (2010) Multiple colonizations from Madagascar and converged acquisition of dioecy in the Mascarene Dombeyoideae (Malvaceae) as inferred from chloroplast and nuclear DNA sequence analyses. Ann Bot 106:343–357

Le Péchon T, Sanchez M, Humeau L, Gigord LD, Zhang LB (2013) Vertebrate pollination of the endemic *Trochetia granulata* (Malvaceae) on Réunion. J Trop Ecol 29:353–356

Lord JM (2008) A test for phylogenetic conservatism in plant-pollinator relationships in Australian and New Zealand alpine floras. New Zeal J Bot 46:367–372

Mansion G, Parolly G, Crowl AA, Mavrodiev E, Cellinese N, Oganesian M, Fraunhofer K, Kamari G, Phitos D, Haberle R, Akaydin G, Ikinci N, Raus T, Borsch T (2012) How to handle speciose clades? Mass taxon-sampling as a strategy towards illuminating the natural history of *Campanula* (Campanuloideae). PLoS ONE 7(11):e50076

Martén-Rodríguez S, Fenster CB, Agnarsson I, Skog LE, Zimmer EA (2010) Evolutionary breakdown of pollination specialization in a Caribbean plant radiation. New Phytol 188:403–417

Menezes de Sequeira M, Jardim R, Silva M, Carvalho L (2007) *Musschia isambertoi* M. Seq., R. Jardim, M. Silva and L. Carvalho (Campanulaceae): a new species from the Madeira Archipelago (Portugal). Ann Jardín Bot Madrid 67:135–146

Micheneau C, Fournel J, Pailler T (2006) Bird pollination in an angraecoid orchid on Réunion island (Mascarene Archipelago, Indian ocean). Ann Bot 97:965–974

Mitchell-Thomé RC (1985) Radiometric studies in Macaronesia. Bol Mus Mun Funchal 37(167):52–85

Navarro-Pérez ML, López J, Fernández-Mazuecos M, Rodríguez-Riaño T, Vargas P, Ortega-Olivencia A (2013) The role of birds and insects in pollination shifts of *Scrophularia* (Scrophulariaceae). Mol Phylogenet Evol 69:239–254

Nyhagen DF, Kragelund C, Olesen JM, Jones CG (2001) Insular interactions between lizards and flowers: flower visitation by an endemic Mauritian gecko. J Trop Ecol 17:755–761

Ojeda I, Santos-Guerra A, Jaén-Molina R, Oliva-Tejera F, Caujapé-Castells J, Cronk Q (2012) The origin of bird pollination in Macaronesian *Lotus* (Loteae, Leguminosae). Mol Phylogenet Evol 62:306–318

Ojeda DI, Santos-Guerra A, Oliva-Tejera F, Valido A, Xue X, Marrero A, Caujapé-Castells J, Cronk Q (2013) Bird-pollinated Macaronesian *Lotus* (Leguminosae) evolved within a group of entomophilous ancestors with post-anthesis flower color change. Perspect Plant Ecol Evol Syst 15:193–204

Olesen JM (1985) The Macaronesian bird-flower element and its relation to bird and bee opportunists. Bot J Linn Soc 91:395–414

Olesen JM (2003) Island pollinators. In: Carvalho MAAPD, Costa GP, Jesus JA, Rodrigues DMM (eds) Island ecosystems: conservation and molecular approach. Centre of Biological and Geological Sciences, Funchal, pp 45–86

Olesen JM, Jordano P (2002) Geographic patterns in plant-pollinator mutualistic networks. Ecology 83:2416–2424

Olesen JM, Valido A (2003a) Lizards as pollinators and seed dispersers: an island phenomenon. Trends Ecol Evol 18:177–181

Olesen JM, Valido A (2003b) Bird pollination in Madeira island. Ardeola 50:67–69

Olesen JM, Valido A (2004) Lizards and birds as generalized pollinators and seed dispersers of island plants. In: Ecología insular, Island ecology: recopilación de las ponencias presentadas en el Symposium de Ecología Insular organizado por la Asociación Española de Ecología Terrestre (AEET) celebrado en Santa Cruz de la Palma (Islas Canarias), 18–24 noviembre 2002, Asociación española de ecología terrestre (AEET), pp 229–249

Olesen JM, Eskildsen LI, Venkatasamy S (2002a) Invasion of pollination networks on oceanic islands: importance of invader complexes and endemic super generalists. Divers Distrib 8:181–192

Olesen JM, Valido A, Dupont Y (2002b) Polinización de plantas canarias. El Indiferente 13:18–29

Olesen JM, Alarcón M, Ehlers BK, Aldasoro JJ, Roquet C (2012) Pollination, biogeography and phylogeny of oceanic island bellflowers (Campanulaceae). Perspect Plant Ecol Evol Syst 14:169–182

Ollerton J, Cranmer L, Stelzer RJ, Sullivan S, Chittka L (2009) Bird pollination of Canary island endemic plants. Naturwissenschaften 96:221–232

Ortega-Olivencia A, Rodríguez-Riaño T, Pérez-Bote JL, López J, Mayo C, Valtueña FJ, Navarro-Pérez M (2012) Insects, birds and lizards as pollinators of the largest-flowered *Scrophularia* of Europe and Macaronesia. Ann Bot 109:153–167

Padilla DP, González-Castro A, Nogales M (2012) Significance and extent of secondary seed dispersal by predatory birds on oceanic islands: the case of the Canary archipelago. J Ecol 100:416–427

Pailler T, Warren B, Labat JN (2002) Biologie de la reproduction de *Aloe mayottensis* (Liliaceae), une espèce endémique de l'île Mayotte (Océan Indien). Can J Bot 80:340–348

Peter CI, Johnson SD (2008) Mimics and magnets: the importance of color and ecological facilitation in floral deception. Ecology 89:1583–1595

Plesner S, Holm PM, Wilson JR (2003) ^{40}Ar–^{39}Ar geochronology of Santo Antão, Cape Verde Islands. J Volcanol Geotherm Res 120:103–121

Prebble JM, Cupido CN, Meudt HM, Garnock-Jones PJ (2011) First phylogenetic and biogeographical study of the southern bluebells (*Wahlenbergia*, Campanulaceae). Mol Phylogenet Evol 59:636–648

Rodríguez-Gironés MA, Santamaria L (2004) Why are so many bird flowers red? PLoS Biol 2(1515–1519):e350

Rodríguez-Rodríguez MC, Valido A (2008) Opportunistic nectar-feeding birds are effective pollinators of bird-flowers from Canary islands: experimental evidence from *Isoplexis canariensis* (Scrophulariaceae). Am J Bot 95:1408–1415

Rodríguez-Rodríguez MC, Valido A (2011) Consequences of plant–pollinator and floral–herbivore interactions on the reproductive success of the Canary islands endemic *Canarina canariensis* (Campanulaceae). Am J Bot 98:1465–1474

Rodríguez-Rodríguez MC, Jordano P, Valido A (2013) Quantity and quality components of effectiveness in insular pollinator assemblages. Oecologia 173:179–190

Roquet C, Sáez L, Aldasoro JJ, Susanna A, Alarcón ML, García-Jacas N (2008) Natural delineation, molecular phylogeny and floral evolution in *Campanula*. Syst Bot 33:203–217

Roquet C, Sanmartín I, García-Jacas N, Sáez L, Susanna A, Wikström N, Aldasoro JJ (2009) Reconstructing the history of Campanulaceae with a Bayesian approach to molecular dating and dispersal-variance analyses. Mol Phylogenet Evol 52:575–587

Schaefer H, Moura M, Belo Maciel MG, Silva L, Rumsey FJ, Carine MA (2011) The Linnean shortfall in oceanic island biogeography: a case study in the Azores. J Biogeog 38:1345–1355

Stuessy TF, Foland KA, Sutter JF, Sanders RW, Silva M (1984) Botanical and geological significance of potassium-argon dates from the Juan Fernandez islands. Science 225:49–51

Thomas H, Félicité M, Adolphe P (2008) Une nouvelle espèce d'*Heterochaenia* DC. (Campanulaceae) de l'ile de Réunion. Acta Bot Gallica 155:245–247

Thomson J (2003) When is it mutualism? Am Nat 162:S1–S9

Uys E, Cron GV (2013) Relationships and evolution in the Drakensberg near-endemic genus, *Craterocapsa* (Campanulaceae). S Afr J Bot 86:79–91

Valido A, Dupont YL, Hansen DM (2002) Native birds and insects, and introduced honey bees visiting *Echium wildpretii* (Boraginaceae) in the Canary islands. Acta Oecologica 23:413–419

Valido A, Dupont YL, Olesen JM (2004) Bird–flower interactions in the Macaronesian islands. J Biogeog 31:1945–1953

Vogel S, Westerkamp C, Thiel B, Gessner K (1984) Ornithophily on the Canary islands (Spain). Plant Syst Evol 146:225–248

Wang Q, Zhou SL, Hong DY (2013) Molecular phylogeny of the platycodonoid group (Campanulaceae s. str.) with special reference to the circumscription of *Codonopsis*. Taxon 62:498–504

Warren BH, Bermingham E, Bowie RC, Prys-Jones RP, Thébaud C (2003) Molecular phylogeography reveals island colonization history and diversification of western Indian ocean sunbirds (*Nectarinia*: Nectariniidae). Mol Phylogenet Evol 29:67–85

Welman WG, Cupido CN (2003) Campanulaceae. In: Germishuizen G, Meyer NL (eds) Plants of Southern Africa: an annotated checklist. Strelitzia 14, Pretoria, National Botanical Institute, pp 336–346

Welsford MR, Johnson SD (2012) Solitary and social bees as pollinators of *Wahlenbergia* (Campanulaceae): single-visit effectiveness, overnight sheltering and responses to flower colour. Arthropod Plant Interact 6:1–14

Yamagiwa J, Basabose AK, Kaleme K, Yumoto T (2005) Diet of Grauer's gorillas in the montane forest of Kahuzi, Democratic Republic of Congo. Int J Primatol 26:345–1373

Yamagiwa J, Basabose AK (2006) Diet and seasonal changes in sympatric gorillas and chimpanzees at Kahuzi-Biega National Park. Primates 47:74–90

Chapter 15
In Search of Phylogeographic Patterns in the Northeastern Atlantic and Adjacent Seas

Sara M. Francisco, Joana I. Robalo, André Levy and Vítor C. Almada

Abstract We reviewed 54 studies on teleost fishes and crustaceans inhabiting European waters to test for the emergence of phylogeographic patterns. Concerning latitudinal variation of genetic diversity, we found that: (1) contrary to the predictions of the "central-margin hypothesis," only a minority of species ($\sim$10 %) revealed higher genetic diversity in the center of their distribution; (2) approximately a third of the fish had a peak of genetic diversity at their southern limit; (3) another substantial fraction of species (41 % for fishes and 72 % for crustaceans) showed little or no latitudinal variation of genetic diversity. Genetic structure expressed by significant FSTs varied widely among species from cases where Atlantic, North Sea, and the Mediterranean seem to correspond to distinct populations, to others where no structure could be detected across their entire range. Given the heterogeneity in sampling schemes we suggest that regular sampling across entire species ranges can improve our understanding of the marine phylogeography in Europe.

15.1 Introduction

The study of marine phylogeography in European shores is now *ca.* 15 years old (Magoulas et al. 1996; Borsa et al. 1997). Data accumulated and, as sequencing became more accessible, more labs were involved and more species were studied. During this period, techniques changed radically and while the initial emphasis was on enzyme electrophoresis it quickly moved to the analyses of mitochondrial DNA and increasingly incorporated nuclear markers, including microsatellites and introns (Sotelo et al. 2009; Almada et al. 2012). Analytical tools also changed and,

S. M. Francisco (✉) · J. I. Robalo · A. Levy · V. C. Almada
Eco-Ethology Research Unit, ISPA University Institute, Rua Jardim do Tabaco 34, 1149-041 Lisbon, Portugal
e-mail: sara_francisco@ispa.pt

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_15,

in particular, estimation of the time to the most recent common ancestor (TMRCA) and of past demography of populations gained accuracy. At the same time, some initial assumptions were reconsidered, as many tools relied on concepts that seem unrealistic for populations subjected to a succession of climatic oscillations during the entire Pleistocene. Thus, models and statistical tools that assume that populations were in mutation-drift equilibrium, or those that assume that populations show a continuous type of growth, are being critically evaluated (Neigel 2002; Kuhner 2009).

The scope of geographical sampling also widened. As the cost of sequencing decreased, more individuals in more populations were included in phylogeographic studies of an increasing number of species, allowing for more fine-scale sampling and the study of local barriers, for the emergence of comparative phylogeography across diverse taxonomic groups, and for the detection of exceptions to broad patterns. For instance, for many years the Atlantic-Mediterranean seaway was thought to be a potential barrier for gene flow. However, the Atlantic-Mediterranean phylogeographic barrier proved to be effective only for some species, while others, often closely related, apparently cross it without restriction (for a review see Patarnello et al. 2007; Kettle et al. 2011).

The need for wide sampling coverage of populations and of comparing different species is particularly acute for the study of the Northeastern Atlantic: a transitional region between the tropics and boreal regions, whose climate has been very dynamic since the Pleistocene. While to a great extent it harbors warm and cold temperate species, it is also the northern limit of some tropical species and the southern limit of some boreal species, thus constituting both a central and marginal habitat (Almada et al. 2013). Furthermore, the climate in this region has changed considerably and often over the last glacial cycles, and during glaciations population ranges are thought to have been driven south or persisted in northern refugia. Evidence is also accumulating that many boreal and cold temperate species survived in peri-glacial refugia (for a review see Maggs et al. 2008).

The central-margin hypothesis (Eckert et al. 2008) assumes the center of a species distribution has a high and stable effective population size (N_e) and a high rate of gene flow (m). Thus, central populations (in this instance, southern populations) should exhibit the highest genetic diversity and harbor the overall most frequent haplotypes. In contrast, marginal (northern) populations should exhibit lower genetic diversity and higher genetic differentiation, and harbor private alleles. The pattern among populations in the Eastern North Atlantic, however, with its history of shifting ranges and patterns of populations connectivity (including presence of refugia), and fluctuations in local populations sizes is likely to be more complex and varied than expected by the central-margin hypothesis.

Maggs et al. (2008) examined several population models, with varying degrees of connectivity, for species distributed along a north–south gradient and expected haplotype networks. Their review of eight benthic species revealed a variety of patterns, and indicated that for some species the admixture of northern refugia populations may lead to the unexpected pattern of greater haplotype diversity in more northern populations. In addition, their coalescent simulations of haplotype

networks for a history of ancestral panmixia followed by geographic isolation call attention to the stochastic nature of genealogies. Overall, their examination indicates that genetic signatures, such as latitudinal patterns of genetic diversity, should be interpreted cautiously and that for a full understanding of population history several aspects should be considered in conjunction—haplotype diversity, monophyly, location of most frequent haplotype and of private haplotypes, rates of migration and identification of admixture events—and in conjunction with the coalescent simulation of different historical scenarios.

In this short review, we compiled information of 54 studies on teleost fishes and crustaceans inhabiting European waters, characterized their sampling, and summarized their phylogeography, in order to identify overall patterns.

15.2 Methods

To compile the set of chapters used in this mini-review, we searched ISI Web of Knowledge database[1] with the following keywords: phylogeograph*, northeastern Atlantic, fish, crustaceans, population structure*, genet* diversity. In addition, we included works that matched our criteria published in two representative journals (*Molecular Ecology* and *Molecular Phylogenetics and Evolution*) between 1997 and 2012. For the complete list of chapters analyzed see Table 15.1. For each chapter, the following data were recorded: distribution area of the species, sampling area, genetic diversity, population structure, time of coalescence estimated for the species, age of the populations, and proposed glacial refugia. When information on any of these items was not reported in a given publication, it was recorded as not available.

The species distributions were retrieved from Fishbase[2] (for fishes) and WoRMS[3] (for crustaceans) and recorded as presence/absence in 13 geographical areas: Arctic, Baltic Sea, North Sea, UK Atlantic coasts, Bay of Biscay, Western Iberian Peninsula, Southern Iberian Peninsula, northwestern African coast, western Mediterranean, eastern Mediterranean, Azores, Madeira and Canaries (Fig. 15.1). The geographic coverage of each study was expressed as the fraction of areas where samples were taken over the total number of areas where the species occurs. An important point in our study was to compare the levels of genetic diversity between northern and southern limits of each species. The southern limit of the species distribution was considered sampled if the geographical area that contains it was sampled in any point (likewise for the northern limit). For the purpose of

[1] www.webofknowledge.com/.

[2] Fish Base (ed, by Froese R, Pauly D). Digital resource available at www.fishbase.org.

[3] WoRMS—World Register of Marine Species (eds by Appeltans W, Bouchet P, Boxshall GA, De Broyer C, de Voogd NJ, Gordon DP, Hoeksema BW, Horton T, Kennedy M, Mees J, Poore GCB, Read G, Stöhr S, Walter TC, Costello MJ). Digital resource available at www.marinespecies.org.

Table 15.1 List of species and publications considered in the present mini-review

	Species	Publication
Fishes	*Alosa alosa*	Faria et al. (2012)
	Alosa falax	Faria et al. (2012)
	Anguilla anguilla	Daeman et al. (2001)
	Aphanopus carbo	Stefanni and Knutsen (2007)
	Atherina presbyter	Francisco et al. (2009)
	Chromis chromis	Domingues et al. (2005)
	Chromis limbata	Domingues et al. (2006b)
	Ciliata mustela	Robalo et al. (2014)
	Conger conger	Correia et al. (2012)
	Coris julis	Aurelle et al. (2003)
	Coryphoblennius galerita	Domingues et al. (2007a)
	Dentex dentex	Bargelloni et al. (2003)
	Dicentrarchus labrax	Lemaire et al. (2005)
	Diplodus puntazzo	Bargelloni et al. (2005)
	Diplodus sargus	Domingues et al. (2007b)
	Engraulis encrasicolus	Magoulas et al. (2006)
	Gasterosteus aculeatus	Mäkinen and Merilä (2008)
	Halobatrachus didactylus	Robalo et al. (2013)
	Helicolenus dactylopterus	Aboim et al. (2005)
	Lipophrys pholis	Francisco et al. (2011)
	Lithognathus mormyrus	Bargelloni et al. (2003)
	Lophius budegassa	Charrier et al. (2006)
	Lophius piscatorius	Charrier et al. (2006)
	Merluccius merluccius	Lundy et al. (1999)
	Mullus surmuletus	Gallarza et al. (2009)
	Pagellus bogaraveo	Bargelloni et al. (2003)
	Pagrus pagrus	Bargelloni et al. (2003)
	Parablennius parvicornis	Domingues et al. (2008b)
	Parablennius sanguinulentus	Domingues et al. (2008b)
	Pholis gunnellus	Hickerson and Cunningham (2006)
	Platichthys flesus	Borsa et al. (1997)
	Pleuronectes platessa	Was et al. (2010)
	Pomatoschistus microps	Gysels et al. (2004)
	Pomatoschistus minutus	Larmuseau et al. (2009)
	Salaria pavo	Almada et al. (2009)
	Salmo salar	Consuegra et al. (2002)
	Sardina pilchardus	Atarhouch et al. (2006)
	Scomber scombrus	Nesbo et al. (2000)
	Scophthalmus maximus	Nielsen et al. (2004)
	Solea solea	Rolland et al. (2007)
	Spondyliosoma cantharus	Bargelloni et al. (2003)
	Sprattus sprattus	Debes et al. (2008)
	Symphodus melops	Robalo et al. (2012)
	Taurulus bubalis	Almada et al. (2012)

(continued)

Table 15.1 (continued)

	Species	Publication
	Thalasoma pavo	Domingues et al. (2008a)
	Thunnus thynnus	Bremer et al. (2005)
	Trachurus trachurus	Karaiskou et al. (2004)
	Tripterygion delaisi	Domingues et al. (2006a)
	Xiphias gladius	Bremer et al. (2005)
Crustaceans	*Calanus helgolandicus*	Papadopoulos et al. (2005)
	Carcinus maenas	Roman and Palumbi (2004)
	Chthamalus montagui	Schemesch et al. (2009)
	Chthamalus stellatus	Schemesch et al. (2009)
	Crangon crangon	Luttikhuizen et al. (2008)
	Euraphia depressa	Schemesch et al. (2009)
	Gammarus duebeni	Rock et al. (2007)
	Homarus gammarus	Triantafyllidis et al. (2005)
	Idotea balthica	Wares and Cunningham (2001)
	Liocarcinus depurator	García-Merchán et al. (2012)
	Macropipus tuberculatus	García-Merchán et al. (2012)
	Maja brachydactyla	Sotelo et al. (2008)
	Meganyctiphanes norvegica	Papetti et al. (2005)
	Mesopodopsis slabberi	Remerie et al. (2006)
	Munida intermedia	García-Merchán et al. (2012)
	Necora puber	Sotelo et al. (2009)
	Nephrops norvegicus	Stamatis et al. (2004)
	Pagurus alatus	García-Merchán et al. (2012)
	Pagurus excavatus	García-Merchán et al. (2012)
	Palinurus elephas	Palero et al. (2008)
	Parapenaeus longirostris	García-Merchán et al. (2012)
	Plesionika heterocamrpus	García-Merchán et al. (2012)
	Pollicipes pollicipes	Quinteiro et al. (2007)

identifying latitudinal patterns, remaining areas were considered central populations. Nei's gene diversity (Nei 1987) was selected to represent genetic diversity. The population structure of a given species was assessed by analysis of molecular variance (AMOVA, Excoffier et al. 1992) and pairwise F_{ST} between locations. We considered that genetic discontinuity existed among locations when the respective pairwise F_{ST} was significant ($p < 0.05$). Due to the potential barriers to gene flow between the northeastern Atlantic and the Mediterranean, and the northeastern Atlantic and the North Sea, articles were searched specifically for the presence of significant F_{ST} involving these three seas. This task was facilitated by the fact that many studies included the assessment of these barriers as an explicit aim. Proposed marine refugia were recorded as: (1) Azores, Canaries, and northwest Africa; (2) southwestern Iberian Peninsula; (3) Mediterranean Sea; (4) western English Channel; (5) southwest Ireland; (6) Iceland and Faroe Islands; and (7) northern Norway (following Maggs et al. 2008). Statistical analyses were performed with

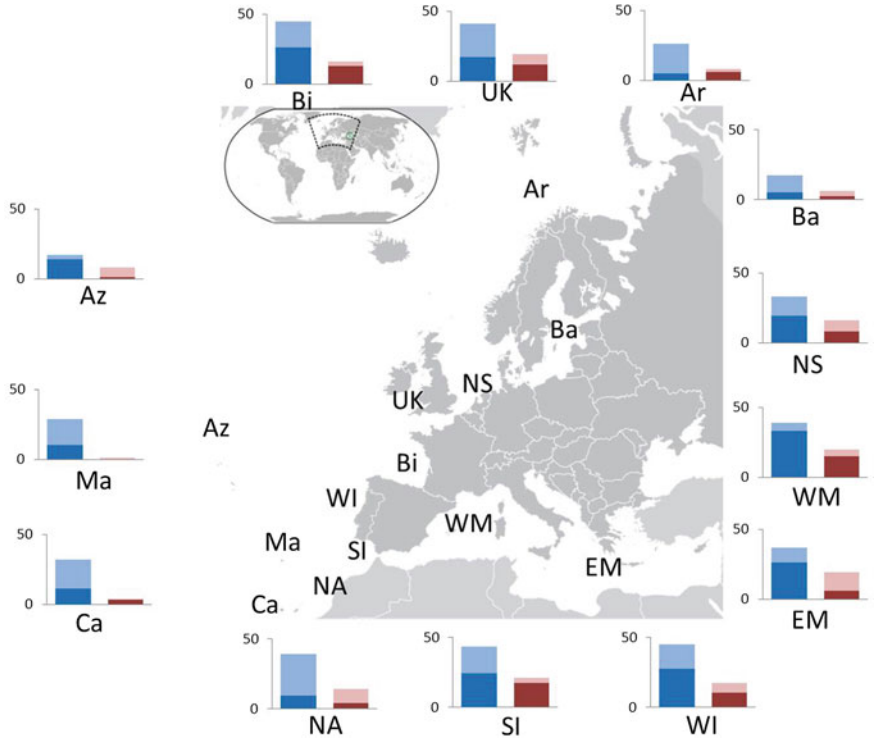


Fig. 15.1 Number of studied species of fish (*blue*) and crustaceans (*red*) present in the area (total height of the bars) and sampled (*dark shade*): Ar—Arctic, Ba—Baltic Sea, NS—North Sea, UK—UK Atlantic coasts, Bi—Bay of Biscay, WI—Western Iberian Peninsula, SI—Southern Iberian Peninsula, NA—Northwestern African coast, WM—Western Mediterranean, EM—Eastern Mediterranean, Az—Azores, Ma—Madeira, Ca—Canaries

the software STATISTICA (StatSoft 2003). Because areas in the Atlantic were less affected by the glaciations than the North Sea and Baltic, we compared gene diversity between these two areas.

15.3 Results

One major feature that emerges from this study is the existence of large gaps and differences in analyses conducted across studies. Very few works sampled the entire distribution area of the species (23 %). An important percentage of the chapters did not assess the age of the populations (74 %). On the positive side, the majority of studies evaluated the existence of population structure for the studied area (93 %). The average geographic coverage found was 54.26 % (S.D. 22.65 %, minimum 15.38 %, maximum 100 %).

15.3.1 Fishes

A total of 39 chapters were analyzed for teleost fishes (Actinopterygii), comprising 50 species belonging to 29 different families and involving 10 different molecular markers.

For the Actinopterygii, only 20 % of the works analyzed in the present review covered the entire area of distribution of the species (Fig. 15.1). In the remaining phylogeographic studies considerable parts of the species range were not sampled. In 34 % of the chapters the sampling was focused at the center of the species' distribution. The northern limit of the species was less sampled than the southern limit of their distribution (30 vs. 56 %) (Fig. 15.2). The sampling coverage was more deficient for the peripheries (Artic 19 %, Baltic 19 %, NW African coast 23 %, Madeira 34 % and the Canaries 34 %), with the exception of the Azores (82 % of the species studied) (Fig. 15.1).

The presence of population structure was not evaluated in only 8 % of the species. Considering the remaining works, 67 % of the species presented genetic structure in the sampled area (Fig. 15.3). Two-thirds of the species sampled in the European Atlantic and the Mediterranean exhibited population structure. Half of the species sampled in the Atlantic revealed genetic differentiation between temperate and North Sea locations. Concerning the Atlantic islands, the following results were found for population structure: 46 % of the studies involving Macaronesian samples found no population structure; 31 % found population differentiation between Azores and European coastal areas; and 23 % showed structure within the Macaronesian Islands.

Genetic diversity was never higher in the North Sea than in the Atlantic, while the opposite pattern (higher diversity in the south) was relatively common (36 %). The pattern of higher diversity in the center of species distributions occurred only in a minority of cases (9 %). Some species appear to have sufficient dispersal and migration (past or present) so that no difference in the distribution of genetic diversity was found along the whole sampled area (41 %) (Fig. 15.4).

Only 20 % of the studies accessed the age of the populations. In 70 % of the species analyzed one or more populations sampled were dated after the Last Glacial Maximum (LGM—18 kya). All studies estimated the age of one or more populations as dating from the last glaciation, before the LGM (120–18 kya).

15.3.2 Crustaceans

A total of 15 chapters were analyzed for the Crustaceans, comprising 23 species (18 from the class Malacostraca and five from the class Maxillopoda), belonging to 16 different families and involving six molecular markers.

Only 30 % of the studies analyzed covered the entire species distribution area (Fig. 15.2). Thirty percent focused on the center of the species distribution, and the

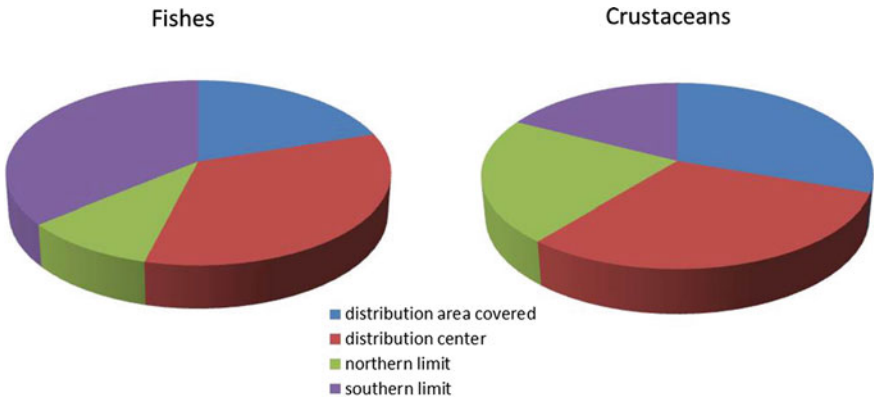


Fig. 15.2 Sampling schemes of the publications considered in this review. *Blue*—sampling covers the entire distribution area of a species; *red*—sampling only covers the center of the species' distribution; *green*—sampling only covers the northern limit and the center of the species' distribution; *purple*—sampling only covers the southern limit and the center of the species' distribution

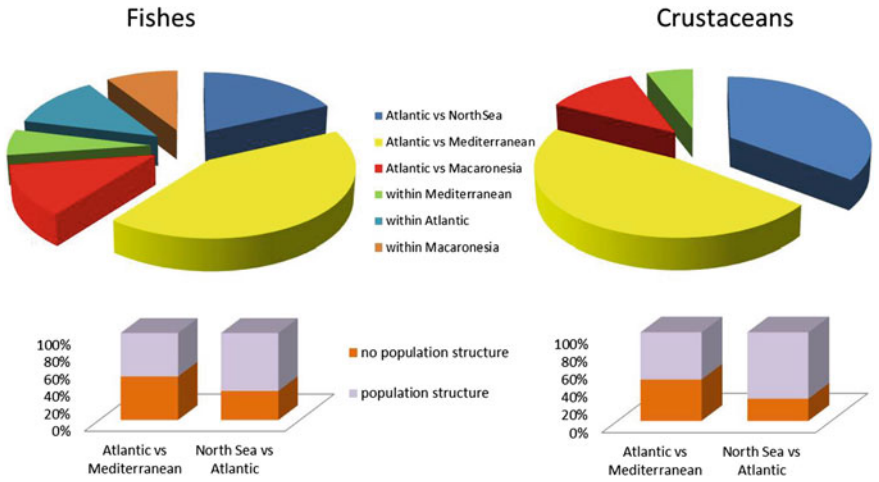


Fig. 15.3 Population structure across the studies compiled for this review: between Atlantic and North Sea (*dark blue*), Atlantic and Mediterranean (*yellow*), Atlantic and Macaronesia (*red*), within the Mediterranean (*green*), within the European Atlantic area (*light blue*) and within Macaronesian islands (*brown*). For the Atlantic versus Mediterranean and North Sea versus Atlantic areas, the proportion of population structure found is shown below the pies

northern limit was slightly more sampled than the south (52 vs. 47 %). The sampling for the crustacean chapters analyzed is more deficient in the peripheries (0 % for Madeira, 13 % for Azores and 33 % for the Baltic Sea) (Fig. 15.1).

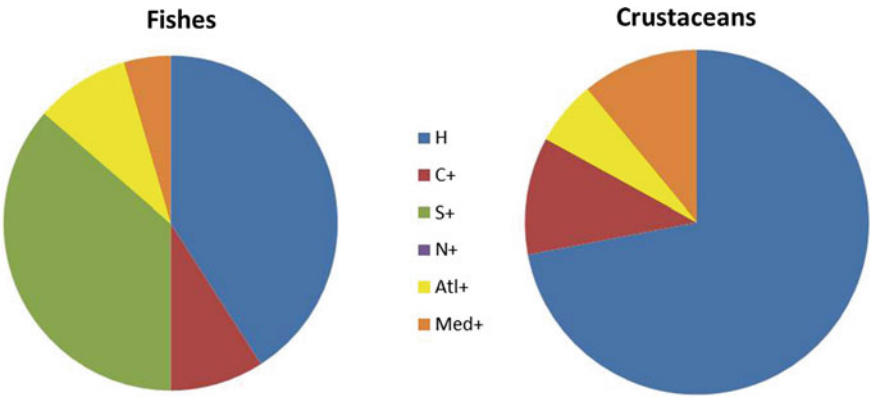


Fig. 15.4 Distribution of the genetic diversity along the sampled areas: H—homogeneous, C+—more diverse in the center of the distribution, S+—more diverse in the southern limit of the distribution, N+—more diverse in the northern limit of the distribution, Atl+—more diverse in the Atlantic than in the Mediterranean, Med+—more diverse in the Mediterranean than in the adjacent Atlantic

The existence of population structure across the studied area was evaluated in 96 % of the studies, with 54 % of the species presenting population genetic structure (52 % of the total of species) (Fig. 15.3). Population structure was found between the Mediterranean and the European Atlantic in 47 % of the studies that addressed this issue, while for the European Atlantic versus North Sea 67 % yielded population structure.

A majority of the studies (72 %) revealed a homogeneous distribution of the genetic diversity across the sampled area (Fig. 15.4). Only 11 % of the species presented a higher genetic diversity in the center of the sampled area. No chapters reported more genetic diversity in the north or south of the species range. Concerning the transition between the Atlantic and the Mediterranean, 6 % of the species presented higher genetic diversity in the Atlantic, while 11 % showed the opposite trend.

The age of the populations was considered in 39 % of the studies, with 44 % of the species with one or more populations dated after the LGM. For 67 % of the works the estimated age of one or more populations was in the last glaciation, before the LGM.

15.3.3 General Patterns

No significant differences in gene diversity were found when comparing the Atlantic coast (Biscay + Iberian Peninsula), North coasts (North Sea + Baltic Sea), and the Mediterranean with Kruskal-Wallis tests (Kruskal and Wallis 1952): total of 73 species: $H = 1.055$, $p = 0.590$, $df = 2$; fishes: $H = 4.018$, $p = 0.134$, $df = 2$; crustaceans: $H = 2.116$, $p = 0.347$, $df = 2$.

Only 21 species were evaluated for their potential glacial refugia. Of these, the Mediterranean may have served as a refugium for five species, namely species that now extend to the adjacent Iberian Atlantic, northern areas for six species and the Iberian Peninsula for nine species.

15.4 Discussion

Several interesting patterns emerge that are supported by both fishes and crustaceans.

1. There are gaps in geographic coverage and substantial heterogeneity in analytical procedures, a situation that reflects the rapid expansion of marine phylogeography and should guide future research in terms of sampling and statistical analysis. The pattern of diversity peaking in the center of the species distribution and decreasing towards the margins holds for only a small minority of species (9 % for fishes and 11 % for crustaceans). We suggest that, in temperate conditions where strong climatic oscillations prevail, the so-called central populations only rarely represent populations that persisted throughout the successive cycles (e.g., Eckert et al. 2008). Rather, many species must have moved up and down during glacial cycles tracking the changes of habitat (Kettle et al. 2011) so that what is now in the center of a distribution will usually reflect the historical contingencies that affected each species.
2. A considerable number of species (30 %) show no structure when different seas are compared (Fig. 15.3). This pattern may have several causes. Perhaps the most likely is that species with high dispersal capabilities and large effective population size may disperse in large numbers, exporting much of the genetic diversity across large geographical scales. This process may erode previous phylogeographic signals (like ancestral polymorphisms) if, after an initial colonization postdating the last glaciation, several millennia of dispersal and mixing elapsed.
3. The observed data underline the importance of the Atlantic as a post-glacial recolonization source for the North Sea. Nevertheless, several studies (e.g., *Carcinus maenas* in Roman and Palumbi 2004; *Taurulus bubalis* in Almada et al. 2012) suggest glacial refugia near the North Sea, and others demonstrate surprisingly long persistence north of the glacial ice for several glacial cycles (e.g., *Pholis gunnellus* in Hickerson and Cunningham 2006). With regard to the Mediterranean, it may have served as a refugium for some species, namely species that now extend to the adjacent Iberian Atlantic (e.g., *Chromis chromis* in Domingues et al. 2005; *Parablennius sanguinolentus* in Domingues et al. 2008b). This rule is, however, far from being universal, as many species may have survived in the Atlantic part of the Iberian Peninsula, along the coast of northwest Africa, and other unglaciated areas in west Europe (see Maggs et al. 2008).

4. The origin of populations (especially in the Atlantic) was dated mainly from the Lower and Middle Pleistocene. Populations with origin estimated after the LGM occur primarily in the North Sea (for cold water species) and Macaronesia, particularly in the Azores (for warm waters species).
5. For thermophilic species, the data seem to support the Azores colonization from Madeira (e.g., Santos et al. 1995). Indeed, when migration was evaluated, more migrants were detected from Madeira to the Azores than the reverse (e.g., *Chromis limbata* in Domingues et al. 2006b). Madeira is, in turn, biogeographically connected with the Canaries, and there are references that point to links between Canaries and Mauritania (e.g., *Tripterygion delaisi* in Domingues et al. 2006a). This colonization route, combined with high sea surface temperatures, probably explains why there are several fishes that are present in the tropics and in the Macaronesian Islands, but not in Europe (e.g., *C. limbata* in Domingues et al. 2006b; *Parablennius parvicornis* in Domingues et al. 2008b).

We are well aware that our sampling is not exhaustive and likely some pertinent literature may have failed to be included in this study. We hope, however, that this survey yielded a representative sample of the scope and patterns of the research being done. Due to lack of time, other important groups could not be included namely mollusks, echinoderms, algae and sea grasses. One major limitation of this survey is that different works often use different molecular markers. There is a substantial probability that slowly evolving markers did not capture the signature of recent events which rapidly evolving ones can retain. The number of studies that we surveyed was not very large, thus separating them by molecular marker would result in a serious loss of information.

Although marine phylogeography is a relatively young science, the evolution of problems and methods of analysis has been very profound. This review included studies with poorly calibrated markers and analytical techniques that differed greatly among studies. A closer communication among laboratories and the building of international projects that cover the entire distribution area of species, with regularly spaced sampling, can have a key role in the evolution of our knowledge. The same holds for quality control of statistical tools and software available, and a better understanding of their underlying assumptions.

Several important projects like Corona,[4] FishPopTrace[5] and a recent MarinERA Project[6] are encouraging examples that mobilized researchers of different countries and did much to share ideas and samples giving rise to a good number of important chapters. We suggest that marine phylogeographers need to advance toward more daring projects. The idea of standardizing sampling patterns that are adequate to represent populations is also a requirement for large scale monitoring of marine communities, while the changes in time of population structure and genetic diversity, in close articulation with oceanographic information, seems

[4] www.biology.duke.edu/corona/.

[5] http://fishpoptrace.jrc.ec.europa.eu/.

[6] http://biocongroup.eu/MarinEra/Welcome.html.

critical to improve our ability to predict the impact of different climate change scenarios. But to achieve this, more long term financing is needed to assure a minimum stability of the networks. So, monitoring at a regular basis seems to us an essential forward step in our field. Finally, a closer cooperation with palaeoclimatologists and palaeoceanographers will add more and more realism to the scenarios we postulate for the glacial conditions, relating the phylogeographic patterns with population models, ocean circulation models and more realistic patterns of larval dispersal.

Acknowledgments This study was funded by the Eco-Ethology Research Unit' Strategic Plan (PEst-OE/MAR/UI0331/2011)—Fundação para a Ciência e a Tecnologia - FCT (partially FEDER funded). SMF was supported by an FCT grant (SFRH/BPD/84923/2012).

We dedicate this chapter in loving memory of Vítor Almada whose tragic loss we suffered during the course of this work.

References

Aboim MA, Menezes GM, Schlitt T, Rogers AD (2005) Genetic structure and history of populations of the deep-sea fish *Helicolenus dactylopterus* (Delaroche 1809) inferred from mtDNA sequence analysis. Mol Ecol 14:1343–1354. doi:10.1111/j.1365-294X.2005.02518.x

Almada V, Almada F, Francisco S et al (2012) Unexpected high genetic diversity at the extreme northern geographic limit of *Taurulus bubalis* (Euphrasen 1786). PLoS ONE 7:e44404

Almada VC, Robalo JI, Levy A et al (2009) Phylogenetic analysis of peri-mediterranean blennies of the genus *Salaria*: Molecular insights on the colonization of freshwaters. Mol Phylogenet Evol 52:424–431. doi:10.1016/j.ympev.2009.03.029

Almada VC, Toledo JF, Brito A et al (2013) Complex origins of the lusitania biogeographic province and northeastern Atlantic fishes. Front Biogeogr 5:20–28

Atarhouch T, Rüber L, Gonzalez EG et al (2006) Signature of an early genetic bottleneck in a population of Moroccan sardines (*Sardina pilchardus*). Mol Phylogenet Evol 39:373–383. doi:10.1016/j.ympev.2005.08.003

Aurelle D, Guillemaud T, Afonso P et al (2003) Genetic study of *Coris julis* (Osteichthyes, Perciformes, Labridae) evolutionary history and dispersal abilities. C R Biol 326:771–785

Bargelloni L, Alarcon JA, Alvarez MC et al (2005) The Atlantic-Mediterranean transition: discordant genetic patterns in two seabream species, *Diplodus puntazzo* (Cetti) and *Diplodus sargus* (L.). Mol Phylogenet Evol 36:523–535. doi:10.1016/j.ympev.2005.04.017

Bargelloni L, Alarcon JA, Magoulas A et al (2003) Discord in the family sparidae (teleostei): divergent phylogeographical patterns across the Atlantic-Mediterranean divide. J Evol Biol 16:1149–1158

Borsa P, Blanquer A, Berrebi P (1997) Genetic structure of the flounders *Platichthys flesus* and *P. stellatus* at different geographic scales. Mar Biol 129:233–246. doi:10.1007/s002270050164

Bremer JRA, Viñas J, Mejuto J et al (2005) Comparative phylogeography of the Atlantic bluefin tuna and swordfish: the combined effects of vicariance, secondary contact, introgression, and population expansion on the regional phylogenies of two highly migratory pelagic fishes. Mol Phylogen Evol 36:169–187. doi:10.1016/j.ympev.2004.12.011

Charrier G, Chenel T, Durand JD et al (2006) Discrepancies in phylogeographical patterns of two European anglerfishes (*Lophius budegassa* and *Lophius piscatorius*). Mol Phylogenet Evol 38:742–754. doi:10.1016/j.ympev.2005.08.002

Consuegra S, De Leaniz CG, Serdio A et al (2002) Mitochondrial DNA variation in Pleistocene and modern Atlantic salmon from the Iberian glacial refugium. Mol Ecol 11:2037–2048

Correia AT, Ramos A a, Barros F et al (2012) Population structure and connectivity of the European conger eel (*Conger conger*) across the north-eastern Atlantic and western mediterranean: integrating molecular and otolith elemental approaches. Mar Biol 159:1509–1525. doi:10.1007/s00227-012-1936-3

Daemen E, Cross T, Ollevier F, Volckaert F (2001) Analysis of the genetic structure of European eel (*Anguilla anguilla*) using microsatellite DNA and mtDNA markers. Mar Biol 139:755–764. doi:10.1007/s002270100616

Debes PV, Zachos FE, Hanel R (2008) Mitochondrial phylogeography of the European sprat (*Sprattus sprattus* L., Clupeidae) reveals isolated climatically vulnerable populations in the mediterranean sea and range expansion in the northeast Atlantic. Mol Ecol 17:3873–3888

Domingues VS, Alexandrou M, Almada VC et al (2008a) Tropical fishes in a temperate sea: evolution of the wrasse *Thalassoma pavo* and the parrotfish *Sparisoma cretense* in the Mediterranean and the adjacent Macaronesian and Cape Verde Archipelagos. Mar Biol 154:465–474. doi:10.1007/s00227-008-0941-z

Domingues VS, Stefanni S, Brito A et al (2008b) Phylogeography and demography of the Blenniid *Parablennius parvicornis* and its sister species *P. sanguinolentus* from the northeastern Atlantic Ocean and the western Mediterranean Sea. Mol Phylogenet Evol 46:397–402. doi:10.1016/j.ympev.2007.05.022

Domingues VS, Almada VC, Santos RS et al (2006a) Phylogeography and evolution of the triplefin *Tripterygion delaisi* (Pisces, Blennioidei). Mar Biol 150:509–519. doi:10.1007/s00227-006-0367-4

Domingues VS, Santos RS, Brito A, Almada VC (2006b) Historical population dynamics and demography of the Eastern Atlantic pomacentrid *Chromis limbata* (Valenciennes 1833). Mol Phylogenet Evol 40:139–147. doi:10.1016/j.ympev.2006.02.009

Domingues VS, Bucciarelli G, Almada VC, Bernardi G (2005) Historical colonization and demography of the Mediterranean damselfish, *Chromis chromis*. Mol Ecol 14:4051–4063. doi:10.1111/j.1365-294X.2005.02723.x

Domingues VS, Faria C, Stefanni S et al (2007a) Genetic divergence in the Atlantic-Mediterranean Montagu's blenny, *Coryphoblennius galerita* (Linnaeus 1758) revealed by molecular and morphological characters. Mol Ecol 16:3592–3605. doi:10.1111/j.1365-294X.2007.03405.x

Domingues VS, Santos RS, Brito A et al (2007b) Mitochondrial and nuclear markers reveal isolation by distance and effects of Pleistocene glaciations in the northeastern Atlantic and Mediterranean populations of the white seabream (*Diplodus sargus*, L.). J Exp Mar Bio Ecol 346:102–113

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. Mol Ecol 17:1170–1188. doi:10.1111/j.1365-294X.2007.03659.x

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Faria R, Weiss S, Alexandrino P (2012) Comparative phylogeography and demographic history of European shads (*Alosa alosa* and *A. fallax*) inferred from mitochondrial DNA. BMC Evol Biol 12:194. doi:10.1186/1471-2148-12-194

Francisco SM, Castilho R, Soares M et al (2009) Phylogeography and demographic history of *Atherina presbyter* (Pisces: Atherinidae) in the North-eastern Atlantic based on mitochondrial DNA. Mar Biol 156:1421–1432. doi:10.1007/s00227-009-1182-5

Francisco SM, Faria C, Lengkeek W et al (2011) Phylogeography of the shanny *Lipophrys pholis* (Pisces: Blenniidae) in the NE Atlantic records signs of major expansion event older than the last glaciation. J Exp Mar Bio Ecol 403:14–20

Galarza JA, Turner GF, Macpherson E, Rico C (2009) Patterns of genetic differentiation between two *co-occurring demersal species: the red mullet (Mullus barbatus) and the striped red mullet (Mullus surmuletus*). Can J Fish Aquat Sci 66:1479–1490

García-Merchán VH, Robainas-Barcia A, Abelló P et al (2012) Phylogeographic patterns of decapod crustaceans at the Atlantic-Mediterranean transition. Mol Phylogenet Evol 62:664–672. doi:10.1016/j.ympev.2011.11.009

Gysels ES, Hellemans B, Pampoulie C, Volckaert FAM (2004) Phylogeography of the common goby, Pomatoschistus microps, with particular emphasis on the colonization of the Mediterranean and the North Sea. Mol Ecol 13:403–417. doi:10.1046/j.1365-294X.2003.02087.x

Hickerson MJ, Cunningham CW (2006) Nearshore fish (*Pholis gunnellus*) persists across the North Atlantic through multiple glacial episodes. Mol Ecol 15:4095–4107. doi:10.1111/j.1365-294X.2006.03085.x

Karaiskou N, Triantafyllidis A, Triantaphyllidis C (2004) Shallow genetic structure of three species of the genus Trachurus in European waters. Mar Ecol Prog Ser 281:193–205

Kettle AJ, Morales-Muñiz A, Roselló-Izquierdo E et al (2011) Refugia of marine fish in the northeast Atlantic during the last glacial maximum: concordant assessment from archaeozoology and palaeotemperature reconstructions. Clim Past 7:181–201. doi:10.5194/cp-7-181-2011

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47:583–621

Kuhner MK (2009) Coalescent genealogy samplers: windows into population history. Trends Ecol Evol 24:86–93. doi:10.1016/j.tree.2008.09.007

Larmuseau MHD, Van Houdt JKJ, Guelinckx J et al (2009) Distributional and demographic consequences of Pleistocene climate fluctuations for a marine demersal fish in the north-eastern Atlantic. J Biogeogr 36:1138–1151. doi:10.1111/j.1365-2699.2008.02072.x

Lemaire C, Versini J-J, Bonhomme F (2005) Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. J Evol Biol 18:70–80. doi:10.1111/j.1420-9101.2004.00828.x

Lundy CJ, Moran P, Rico C et al (1999) Macrogeographical population differentiation in oceanic environments: a case study of European hake (*Merluccius merluccius*), a commercially important fish. Mol Ecol 8:1889–1898

Luttikhuizen PC, Campos J, van Bleijswijk J et al (2008) Phylogeography of the common shrimp, *Crangon crangon* (L.) across its distribution range. Mol Phylogenet Evol 46:1015–1030. doi:10.1016/j.ympev.2007.11.011

Maggs CA, Castilho R, Foltz D et al (2008) Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. Ecology 89:S108–S122

Magoulas A, Castilho R, Caetano S et al (2006) Mitochondrial DNA reveals a mosaic pattern of phylogeographical structure in Atlantic and Mediterranean populations of anchovy (*Engraulis encrasicolus*). Mol Phylogenet Evol 39:734–746. doi:10.1016/j.ympev.2006.01.016

Magoulas A, Tsimenides N, Zouros E (1996) Mitochondrial DNA phylogeny and the reconstruction of the population history of a species: the case of the European anchovy (*Engraulis encrasicolus*). Mol Biol Evol 13:178–190

Mäkinen HS, Merilä J (2008) Mitochondrial DNA phylogeography of the three-spined stickleback (*Gasterosteus aculeatus*) in Europe-evidence for multiple glacial refugia. Mol Phylogenet Evol 46:167–182. doi:10.1016/j.ympev.2007.06.011

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Neigel JE (2002) Is FST obsolete? Conservation Genetics 3:167–173. doi:10.1023/A:1015213626922

Nesbø CL, Rueness EK, Iversen S a et al (2000) Phylogeography and population history of Atlantic mackerel (*Scomber scombrus* L.): a genealogical approach reveals genetic structuring among the eastern Atlantic stocks. Proc Biol Sci 267:281–292. doi:10.1098/rspb.2000.0998

Nielsen EE, Nielsen PH, Meldrup D, Hansen MM (2004) Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. Mol Ecol 13:585–595

Palero F, Abelló P, Macpherson E et al (2008) Phylogeography of the European spiny lobster (*Palinurus elephas*): Influence of current oceanographical features and historical processes. Mol Phylogenet Evol 48:708–717. doi:10.1016/j.ympev.2008.04.022

Papadopoulos LN, Peijnenburg KTCA, Luttikhuizen PC (2005) Phylogeography of the calanoid copepods *Calanus helgolandicus* and *C. euxinus* suggests Pleistocene divergences between Atlantic, Mediterranean, and black sea populations. Mar Biol 147:1353–1365. doi:10.1007/s00227-005-0038-x

Papetti C, Zane L, Bortolotto E et al (2005) Genetic differentiation and local temporal stability of population structure in the euphausiid *Meganyctiphanes norvegica*. Mar Ecol Prog Ser 289:225–235. doi:10.3354/meps289225

Patarnello T, Volckaert FMJ, Castilho R (2007) Pillars of Hercules: is the Atlantic-Mediterranean transition a phylogeographical break? Mol Ecol 16:4426–4444. doi:10.1111/j.1365-294X.2007.03477.x

Quinteiro J, Rodríguez-Castro J, Rey-Méndez M (2007) Population genetic structure of the stalked barnacle *Pollicipes pollicipes* (Gmelin 1789) in the northeastern Atlantic: influence of coastal currents and mesoscale hydrographic structures. Mar Biol 153:47–60. doi:10.1007/s00227-007-0783-0

Remerie T, Bourgois T, Peelaers D et al (2006) Phylogeographic patterns of the mysid *Mesopodopsis slabberi* (Crustacea, Mysida) in Western Europe: evidence for high molecular diversity and cryptic speciation. Mar Biol 149:465–481. doi:10.1007/s00227-005-0235-7

Robalo JI, Castilho R, Francisco SM et al (2012) Northern refugia and recent expansion in the North Sea: the case of the wrasse *Symphodus melops* (Linnaeus 1758). Ecol Evol 2:153–164. doi:10.1002/ece3.77

Robalo JI, Crespo AM, Castilho R et al (2013) Are local extinctions and recolonizations continuing at the colder limits of marine fish distributions? *Halobatrachus didactylus* (Bloch and Schneider 1801), a possible candidate. Mar Biol 160:2461–2467. doi:10.1007/s00227-013-2241-5

Robalo JI, Lima CS, Francisco SM, et al (2014) Phylogeography of the fivebeard rockling (*Ciliata mustela*, Linnaeus 1758) J Phylogen Evol Biol 2:123. doi: 10.4172/2329-9002.1000123

Rock J, Ironside J, Potter T et al (2007) Phylogeography and environmental diversification of a highly adaptable marine amphipod, *Gammarus duebeni*. Hered (Edinb) 99:102–111. doi:10.1038/sj.hdy.6800971

Rolland JL, Bonhomme F, Lagardère F et al (2007) Population structure of the common sole (*Solea solea*) in the Northeastern Atlantic and the Mediterranean Sea: revisiting the divide with EPIC markers. Mar Biol 151:327–341. doi:10.1007/s00227-006-0484-0

Roman J, Palumbi SR (2004) A global invader at home: population structure of the green crab, *Carcinus maenas*, in Europe. Mol Ecol 13:2891–2898. doi:10.1111/j.1365-294X.2004.02255.x

Santos RS, Hawkins S, Monteiro LR et al (1995) Marine research, resources and conservation in the Azores. Aquat Conserv Mar Freshw Ecosyst 5:311–354. doi:10.1002/aqc.3270050406

Shemesh E, Huchon D, Simon-Blecher N, Achituv Y (2009) The distribution and molecular diversity of the Eastern Atlantic and Mediterranean chthamalids (Crustacea, Cirripedia). Zool Scr 38:365–378. doi:10.1111/j.1463-6409.2008.00384.x

Sotelo G, Morán P, Posada D (2008) Genetic Identification of the Northeastern Atlantic Spiny Spider Crab as *Maja Brachydactyla* Balss 1922. J Crustac Biol 28:76–81. doi:10.1651/07-2875R.1

Sotelo G, Posada D, Morán P (2009) Low-mitochondrial diversity and lack of structure in the velvet swimming crab *Necora puber* along the Galician coast. Mar Biol 156:1039–1048. doi:10.1007/s00227-009-1148-7

Stamatis C, Triantafyllidis A, Moutou KA, Mamuris Z (2004) Mitochondrial DNA variation in Northeast Atlantic and Mediterranean populations of Norway lobster, *Nephrops norvegicus*. Mol Ecol 13:1377–1390. doi:10.1111/j.1365-294X.2004.02165.x

StatSoft I (2003) STATISTICA (data analysis software system), version 6. www.statsoft.com

Stefanni S, Knutsen H (2007) Phylogeography and demographic history of the deep-sea fish *Aphanopus carbo* (Lowe, 1839) in the NE Atlantic: vicariance followed by secondary contact or speciation? Mol Phylogenet Evol 42:38–46. doi:10.1016/j.ympev.2006.05.035

Triantafyllidis A, Apostolidis AP, Katsares V et al (2005) Mitochondrial DNA variation in the European lobster (*Homarus gammarus*) throughout the range. Mar Biol 146:223–235. doi:10.1007/s00227-004-1435-2

Wares JP, Cunningham CW (2001) Phylogeography and historical ecology of the North Atlantic intertidal. Evolution 55:2455–2469

Was A, Gosling E, Hoarau G (2010) Microsatellite analysis of plaice (*Pleuronectes platessa* L.) in the NE Atlantic: weak genetic structuring in a milieu of high gene flow. Mar Biol 157:447–462. doi:10.1007/s00227-009-1331-x

Chapter 16
The Evolutionary Space Model to be Used for the Metagenomic Analysis of Molecular and Adaptive Evolution in the Bacterial Communities

E. V. Pershina, A. S. Dolnik, G. S. Tamazyan, K. V. Vyatkina, Y. B. Porozov, A. G. Pinaev, S. O. Karimov, N. A. Provorov and E. E. Andronov

Abstract Recent progress in metagenomics resulted in rapid accumulation of data on genetic diversity of microorganisms. This diversity is mostly represented by uncultured microorganisms, never described in regard to phenotype. Therefore, former phenotypic classification system of bacteria came out to be inapplicable to metagenomics and was thus replaced with a genotypic system, built upon the 16S rRNA gene16S rRNA gene. Metagenomics operates with nucleotide sequences instead of species. This shift in biodiversity assessment required a new classification system. In this study, we attempted to develop such a system. We call it the

E. V. Pershina · A. G. Pinaev · N. A. Provorov · E. E. Andronov (✉)
All-Russia Research Institute for Agricultural Microbiology (ARRIAM) RAAS,
St. Petersburg-Pushkin, Podbelskogo Highway, 3, Petersburg, Russia 196608
e-mail: eeandr@gmail.com

E. V. Pershina
e-mail: pershina.elizaveta@yandex.ru

A. S. Dolnik
Resource Center of the Development of Molecular and Cellular Technologies,
St. Petersburg State University, St. Petersburg, University Embankment, 7-9,
Petersburg, Russia 199034
e-mail: alexander.dolnik@gmail.com

G. S. Tamazyan · S. O. Karimov
St. Petersburg State University, St. Petersburg, University Embankment, 7-9,
Petersburg, Russia 199034

K. V. Vyatkina
Education and Research Center of Nanotechnology RAS, St. Petersburg,
Hlopyna st., 8, Block 3, Petersburg, Russia 195220
e-mail: kira@math.spbu.ru

Y. B. Porozov
St. Petersburg National University of Information Technologies Mechanics and Optics,
St. Petersburg, Kronverkskiy Avenue, 49, Petersburg, Russia 197101
e-mail: porozov@ifc.cnr.it

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_16,

Evolutionary Space (ES). ES is a metric multidimensional space, where each point represents a single 16S rRNA sequence. These points are geometrically spaced according to the genetic distances between corresponding sequences. ES is aimed to represent genetic variability of all currently existing and theoretically predicted 16S rRNA genes and to identify obscure evolutionary patterns defining this enormous biodiversity. We used the mathematics of regular simplexes to identify dimensional properties of ES, constructed upon the distance matrix, derived from sequences stored in the SILVA database. It appeared to be 13D. After mapping the 16S rRNA database in ES a set of evolutionary patterns were observed on a series of ES slices (e.g., evolutionary hollows likely to represent probable positions of ancestral genes). Different bacterial phyla formed well-defined areas within the space. ES can also be used in the practice of metagenomic analysis (e.g., in the analysis of microbial communities' succession). To show this we examined the dynamics of two microbiomes, exposed to salt stress in natural and artificial conditions. ES helps us to study these dissimilar microbiomes as mathematical objects possessing several geometric characteristics (e.g., shape, density, trajectory, and vector of community development). Although some essential biological questions could benefit from ES, its practical implementation requires collaboration in many fields of natural science, including bioinformatics, mathematics and, to some extent, astrophysics.

16.1 Introduction

Recent developments in high-throughput sequencing technologies resulted in close to limitless possibilities to study microbiomes' genetic diversity (Roesch et al. 2007; Bartram et al. 2011; Simon and Daniel 2011; Kim et al. 2013). Metagenomics expansion revealed the enormous prokaryotic diversity ($\sim 10^9$ species), far exceeding that of macroorganisms (Dykhuizen 1998; Hughes et al. 2001). This diversity is mostly represented by phantom microorganisms, with phenotypes one can hardly imagine. This "uncultured biosphere" brought conventional microbial systematics and phylogeny to chaos, thus left us pending for a modern and effective system of prokaryotic taxonomy (Lilburn and Garrity 2004; Pace 2009). In its turn, the lack of systematic convenience forced metagenomics to work with nucleotide sequences (e.g., 16S rRNA genes16S rRNA gene) instead of defined species. During the metagenomic analysis 16S rRNA genes16S rRNA gene are usually grouped together using formal sequence similarity criteria resulting in the so-called Operational Taxonomic Units (OTUs) (usually OTUs gained at 97 % rDNA identity are considered as species equivalent). This allows operating with all sorts of sequences, whether named or not (Hao et al. 2011). Since metagenomics mostly deals with sequences it entirely depends on time-consuming alignment procedures. This step makes it difficult to include new sequences into analysis, since every time the entire procedure has to be repeated. In some cases it seriously hampers the comparative analysis of large datasets (DeSantis et al. 2006). Thus, despite many software packages are available (Rudi et al. 2006; Cardenas et al. 2009; Pommier

et al. 2009; Schloss et al. 2009; Caporaso et al. 2010; Ludwig et al. 2004), there is a need for novel approaches to cope with emerging data on prokaryotic diversity.

It's worth noting the creation of such a system became possible only when sufficient data on the genetic diversity of microorganisms were accumulated. The emergence of Woese's concept (Woese 1987) made 16S rRNA gene16S rRNA gene extraordinary popular in microbiology. Thus, now 16S rRNA gene16S rRNA gene has the most comprehensive database of all known genes (e.g., curated databases, such as RDP, Ribosomal Database Project (Cole et al. 2009) store nearly 3 million sequences). So today we have obtained the constantly updating map of evolutionary history of prokaryotes. This means we are very close to the construction of a natural system for prokaryotic classification (yet build for a single gene), discussed by many evolutionists ever since Linnaeus. He was the first to counter artificial and natural systems, defining the latter one as a system, which "…by itself indicates even plants that are omitted; which enumeration in a catalogue never does" (Linnaeus 1751). It's hard to find a better definition for modern problems in the prokaryotic taxonomy. Natural system for prokaryotic classification, in turn, must become the basis for the analysis of microbial biodiversity.

Here, we present a sketch of the natural system (called evolutionary space—ES) aimed to put every single 16S rRNA tag in a "system of coordinates" (in this case a set of coordinates can be used as an individual "name"). Not only can this system be applied to "name" undescribed microbes, but to calculate integral parameters, describing the structure and dynamics of microbiomes represented as geometric structures in a multidimensional space.

The basic idea of the ES system is to put 16S rRNA sequences in a metric space, representing them by dots, while geometric distances between these dots correlate with genetic distances (expressed in *p*-distance) between corresponding sequences. This approach is quite similar to classic goal of MDS (Multidimensional Scaling). It's worth noting, the idea of such constructions is not new. Many researchers have demonstrated similar approaches, though their implementations were limited to low-dimensional (2–3D) spaces (Hughes et al. 2004; Lee et al. 2006) or the projections of the multidimensional spaces, imposing restrictions on the amount of simultaneously presented sequence data (Garrity and Lilburn 2002; Kitazoe et al. 2011). The central idea of our study was not to simply divide taxonomic groups of prokaryotes, but to detect evolutionary patterns underlying existing diversity of 16S rRNA gene16S rRNA gene. It turned out that we could actually see these patterns. The only thing we needed was to find the appropriate dimensionality of the corresponding metric space.

16.2 Dimensionality of Evolutionary Space

We often hear people say our system is overloaded with mathematics. In fact, it is not fair. Our constructions have only two aspects to some extent related to mathematics. The first one is the reconstruction of the 16S rRNA sequence evolutionary patterns; the second one is the determination of ES dimensionality.

Intuitively, genetic "distances" between sequences have the meaning of distances (although it is not always so *in stricto sensu*). Particularly, widespread phylogenetic trees are also based on this statement—they essentially are 2D interpretations of the distance matrix data. Obviously, this "planar" interpretation of initially multidimensional distance data leads to a serious loss of information. Figuratively, we cannot see the forest for the trees. By forest we mean different patterns of the 16S rRNA evolution, which can only be seen if all distances in the matrix were reflected. This well-known phenomenon has been repeatedly cited in literature (Huson and Bryant 2006; Lee et al. 2006; Brandes and Cornelsen 2009). To "extract" the "evolutionary message" we have to put corresponding distance matrix data into a multidimensional space, in other words to perform MDS. This task has been posed by a number of researchers, including attempts to use MDS in the classification manual for the last edition of Bergy's manual for systematic bacteriology (Garrity and Lilburn, 2002). Particularly a standard PCA analysis was performed with the bacterial 16S rRNA distance matrix. As a result authors obtained projections of the so-called "taxonomic space". Although they clearly demonstrated various phylogenetic relationships between and within taxa, they failed to map the whole biodiversity and what is most important they failed to describe any kind of evolutionary patterns, since they operated with the low-dimensional projections.

For the correct reconstruction of evolutionary patterns we need to map all 16S rRNAs from the database in the space with sufficient dimensionality. In such a case conventional MDS analysis would require a space of several hundreds, or even a thousand, dimensions. It is virtually impossible to work with such complicated systems. Thus, we tried to display database data in a multidimensional space, choosing its dimensionality according to some intrinsic properties of the database.

For the ES construction we used geometric tools, such as regular simplexes (polyhedrons)—multidimensional generalization of a regular triangle. The dimensionality of a simplex defines the dimensionality of the corresponding geometric space (2D space for regular triangle, 3D space for regular tetrahedron, generally nD space for $(n + 1)$D simplex). Nucleotide sequences with equal pairwise distances can also form regular simplexes (Fig. 16.1).

For example, 10 sequences shown on Fig. 16.1 have equal pairwise distances (2-nucleotide difference between each pair of sequences). In terms of geometry, they form 10 vertices of a simplex, which can only be represented in a space of at least 9 dimensions. Hence, we can determine the dimensionality of ES by fetching out regular simplexes with maximal dimensionality within a processed database.

16.3 Construction of the Evolutionary Space and Arrangement of 16S rRNA Sequences

To find the simplex with maximal number of vertices (which corresponds to the essential dimensionality of ES) we used an appropriate release of SILVA database (Pruesse et al. 2007) for 16S rRNA genes16S rRNA gene SSURef_104_SILVA_

10 sequences

1. TAAAAAAAAA
2. ATAAAAAAAA
3. AATAAAAAAA
4. AAATAAAAAA
5. AAAATAAAAA
6. AAAAATAAAA
7. AAAAAATAAA
8. AAAAAAATAA
9. AAAAAAAATA
10. AAAAAAAAAT

Distance matrix

	1	2	3	4	5	6	7	8	9	10
1										
2	2									
3	2	2								
4	2	2	2							
5	2	2	2	2						
6	2	2	2	2	2					
7	2	2	2	2	2	2				
8	2	2	2	2	2	2	2			
9	2	2	2	2	2	2	2	2		
10	2	2	2	2	2	2	2	2	2	

Fig. 16.1 An example of 10 sequences forming regular simplex

NR_99 (available at http://www.arb-silva.de/no_cache/download/archive/release_104/Exports/). It contained aligned high quality 16S rRNA gene16S rRNA gene sequences with similarity less than 99 % and no less than 1,000 bp in length.

After exclusion of archaea, we obtained 210,651 nucleotide sequences (the length of the alignment was about 30,000 positions) corresponding to the bacterial 16S rRNA genes16S rRNA gene. The distance matrix was created by using *p*-distance (pairwise deletion). The search of simplexes was performed in [0,251–0,269] range of distances. For simplex search the "greedy" algorithm was applied. It is based on the identification of candidate sequences characterized by maximal scores of their pairwise distances stacking within a given range, followed by step-by-step stochastic extension of the list. Calculations resulted in the identification of several sets of simplexes with different sizes. For further analysis we selected 25 simplexes containing maximal number of sequences (14 sequences each, Fig. 16.2).

Simplex 1 EU773611; EU491566; AJ542543; AY485285; AB355037; X86688; EU703430; Y10649; EF096697; EF516823; EU804917; AY571792; AM420109; AJ306801
Simplex 2 EU469976; EU503653; GQ502583; AF189244; AY212563; AY907749; GQ397076; FJ628180; DQ814080; EU669608; DQ811945; AB191897; GQ346956; EU245865
Simplex 3 FJ231137; EU135237; DQ795973; EU776122; AY863081; EU881151; EF020301; EU802835; AB488334; AB300126; EU038002; EU246179; FJ545465; CU924649 116
Simplex 4 AF419696; EU506479; EU507872; DQ811928; D11348; FJ456773; X71862; FN563192; CP001110; FJ648694; AF068427; EU335420; AY743263; FN556062

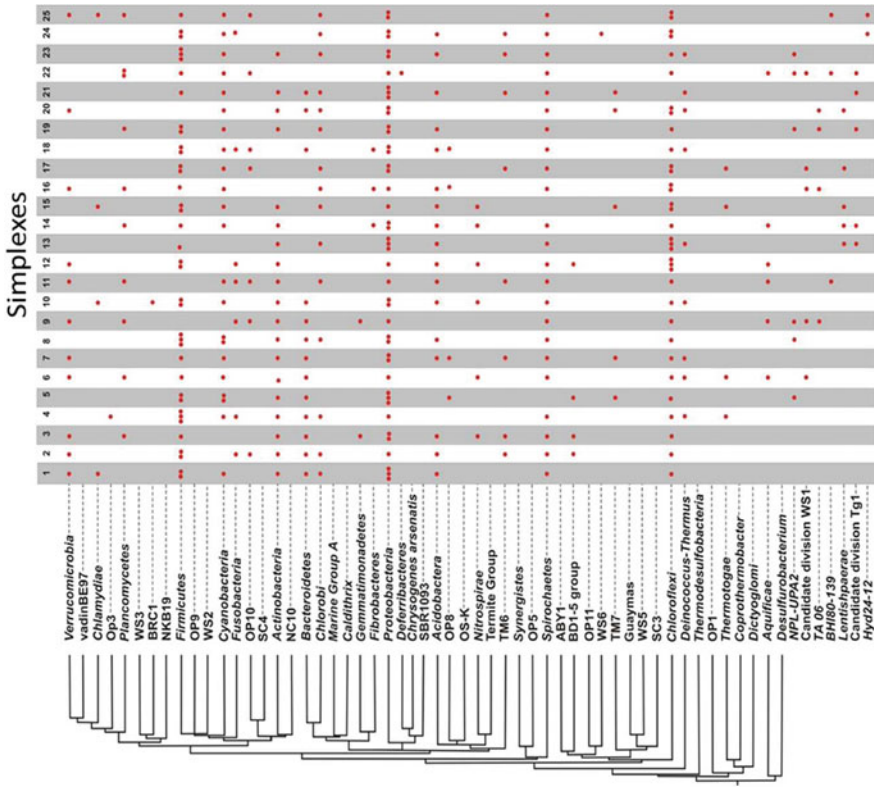


Fig. 16.2 Simplexes found in SILVA database distance matrix and taxonomic assignment (phylum) of their vertices

Simplex 5 EU381735; EF688230; EU370505; EF454921; EU799550; EF575061; FJ821610; FN401325; GU061319; FJ873298; AY280413; EU135375; FJ592895; GQ350871

Simplex 6 FJ438004; GQ246374; FJ881166; EU005687; X73976; EU463251; DQ337095; AY225654; AY605151; FJ478836; FJ628268; FJ901103; CP001080; CU925754

Simplex 7 DQ803694; EU767531; X12742; X81063; EU869405; EU134585; EU360497; AY571796; AY197394; AB177131; EF203193; FJ976270; EU134048; FJ891053

Simplex 8 EU465688; EU511290; FJ366892; AB188635; AY663886; GU127275; EU775151; FJ717259; EU804722; FJ456653; EU491403; AJ431238; CU922689; FJ516821

Simplex 9 EF575007; FJ748813; EU366375; AM712329; DQ248296; FJ983028; GQ263308; FJ802296; AY605160; EF076074; DQ906017; AB294345; CU923425; DQ330595

Simplex 10 EU074225; CT573820; CU925797; DQ800076; EU037954; FJ976253; AB464934; DQ308543; FJ192842; EF019248; EU250258; AB243263; EU133963; X84212
Simplex 11 AB277853; EU478629; EF522262; EU772741; EU635952; AY907749; CP001099; EU134803; EU159562; AB245338; GQ397047; AB192244; CU923893; DQ499300
Simplex 12 EU507587; U32593; M24483; AM712329; FJ826329; AJ867904; EF019021; AF543503; DQ676428; EU266879; CU922282; EU043840; GQ340131; DQ906038
Simplex 13 EU505590; FJ858737; FJ628297; AB240485; EU134568; U91515; EU132320; AB031999; CU921210; EU134128; GQ264185; EU289449; CU920242; FJ625343
Simplex 14 FJ748815; EU503864; FJ159133; EU010170; EF453815; EU135522; EF018434; EF515949; X86774; AB198654; CU918198; AB462555; CR933027; FJ264554
Simplex 15 AY114316; EU463474; GQ275102; FJ382145; AY672075; DQ906842; DQ005880; GQ264171; M79383; EU134919; AM934777; EU134038; CU921544; EU850520
Simplex 16 EF520637; FJ873260; DQ809643; EU617874; AB286524; EU470375; FN430655; FJ478875; DQ811949; EF203193; EU245088; FJ478622; CU925754; EF192905
Simplex 17 AB192054; EU509270; GQ441271; AY188316; DQ906842; AB286350; GU118530; AY945884; AB088905; CU922949; FJ517055; CU918272; CU927871; AY114333
Simplex 18 EU778001; EU763449; AY726960; AB355083; FJ592715; FJ516977; EF190824; AY947962; CP000814; EU132011; FJ712505; EU592424; EU134203; FM873402
Simplex 19 AF317763; EU459226; EU639371; FJ002234; EF592610; EF205470; EU133431; FJ167503; AY913233; AB198604; CU924983; EU662508; FJ712493; AB282966
Simplex 20 EF019165; EU982406; FJ425646; FN563173; DQ383304; AY349381; EU134307; AF093251; CU918643; CU924912; EU133993; EU247889; EU245649; EU885068
Simplex 21 AB192219; D11348; EU802784; CU923009; AJ291826; EU773650; GU061962; EU334768; FJ592772; EU385703; AY571473; FJ825446; AB465709; FJ004754
Simplex 22 AB302409; EU669636; EU434533; FJ493498; FJ790619; FJ746187; EF379616; CU921631; EU915265; AF393378; DQ676384; AB234287; AB525461; AB089051
Simplex 23 AY862537; EF097759; EU775762; GQ487946; AF385521; AJ306807; EU050858; EU802639; AY913288; EU236294; X89045; CU925964; AF521187; AB294345
Simplex 24 DQ015655; EU507714; FJ425597; CP001739; FJ802178; FJ985790; FJ628291; EU409852; GQ402806; EF688228; EU721768; CU926616; DQ988318; GQ249498

Simplex 25 FN554390; EU074225; AY266450; AB034054; FJ002173; EF522341; AJ299413; FJ628241; GQ355003; AF402980; CU927201; EU181504; AB237731; FJ879997

To map database sequences in ES, we used vertices of the simplex number 6. Generally, we used simplex vertices as "GPS satellites", determining the position of each of 210,651 sequences in the 13D metric space on the principle similar to the GPS positioning (Dolnik et al. 2012).

Correlations between pairwise genetic distances and their geometric analogs were calculated using the Mantel test (Mantel and Valand 1970).

To visualize the distribution of points in a 13-dimensional space we constructed several orthogonal 2D planes and made a series of "thin" slices (parallel and orthogonal). For each of the axes we created a plane passing through the geometrical center of the simplex (for orthogonal sections). We obtained one thousand sections in total. The most interesting of them are presented in this publication.

16.4 Simplexes Composition

In Fig. 16.2, you can see main bacterial taxa—on the left hand they are represented as branches of the phylogenetic tree (only major bacterial phyla), on the right—by dots representing the vertices of each of the 25 independent simplexes. All of the vertices within every simplex correspond to the deeply divergent gene variants, belonging to different bacterial phyla (Fig. 16.2). Objects of different phyla were evenly represented within the simplex, except for those of *Spirochaetes* and *Chloroflexi*, which occurrence frequencies were about 7 and 5 times higher than expected, respectively (based on their number in the database). The rest of the phyla distributed according to their relative abundance in the database. Besides the high level of diversity, the *Firmicutes* and the *Proteobacteria* are broadly represented in the database, and therefore occur more than once within the same simplex.

16.5 Diversity of Bacterial Taxa and Evolutionary Patterns of 16S rRNA Gene

Using simplex vertices as a kind of "GPS-satellites" we mapped all 16S rRNA sequences from SILVA release. Before moving to the explanation of evolutionary patterns it should be emphasized that database data are biased and can only partially reflect bacterial biodiversity in the biosphere. Obviously, the composition of modern databases depends entirely on the intensity of research activity in respect to certain taxa (for example, databases are greatly overcrowded with proteobacterial sequences). There were many attempts to improve this mismatch (Bietz and Lee 2009; Giongo et al. 2010; Werner et al. 2012). Hopefully this problem will be

completely solved within the next decade. Nonetheless, we shouldn't forget our biodiversity studies are limited to modern databases.

What are those evolutionary patterns we expect to see while analyzing the 16S rRNA biodiversity? We know that 16S rRNA gene16S rRNA gene have been changing since the beginning of molecular evolution, even during the pre-cellular era. Modern diversity is the direct result of continuous divergent evolution of 16S rRNA gene16S rRNA gene structure. So, is it possible to reconstruct the evolutionary story of 16S rRNA gene16S rRNA gene and make assumptions about the rate (number of nucleotide substitutions for certain time period in different taxa) and directions (polyphyletic or monophyletic) of its evolution?

The evolutionary process can probably be explained in a multidimensional space as a radial and irreversible extension from the initial point (ancestral sequence). This point can be identified geometrically and is probably empty due to the disappearance of ancestral gene variants because of genetic drift, mutational pressure, or microbiological mass extinctions and never can be restored during the evolution, irreversible according to Dollo's law (Dollo 1893). If we map one of the modern 16S rRNA databases in this space, we will probably see filled areas (regions which may correspond to some discrete taxonomic groups) as well as evolutionary hollows (extinct ancestral gene variants and places for some upcoming modifications of 16S rRNA gene16S rRNA gene not yet embodied in the evolutionary history or even places which have never been filled and will never be (e.g., functionally inactive forms of rRNA). The difference between hollows, formed by the latter two variants, however, can only be distinguished from the overall context.

It's quite important to note, the aforementioned evolutionary patterns (e.g., filled and hollow areas) cannot be seen on projections. Therefore, to visualize them we analyzed a series of orthogonal slices.

Figure 16.3 represents several slices of ES. The picture we see comes close to our hypotheses. We can see areas of compact localization of points, painted in different colors (Fig. 16.3), while each color matches a single bacterial phylum. You can clearly see how phyla separate. At the same time we can see empty areas. If we look at some sections with more detail, we can find some interesting patterns (Fig. 16.4) indicating the formation of discrete bacterial taxa during the evolution.

The red star in Fig. 16.4a marks a hypothetical center of the evolutionary space (the probable place of localization of the last common ancestor of bacteria). Assuming the expansion of evolutionary space is radial, we determine this point by building 2 axes passing lengthwise the bacterial phyla elongated in space (like *Cyanobacteria* and *Bacteroidetes*). It is very rough construction of course. As you can see, this place is empty, while main bacterial phyla are spreading from it according to the rate of their divergence from the last common ancestor. This structure has reminded us a candle flame with the *Proteobacteria* (the most evolutionarily young phylum) at very top. Within the *Proteobacteria* we found a place, very similar to the evolutionary hollow (a place for proteobacterial ancestors). Probably, the formation of hollows is peculiar for dynamically evolving phyla.

Besides the general evolutionary scenario some specific patterns were found. Among them are the evolutionary patterns associated with the origination of

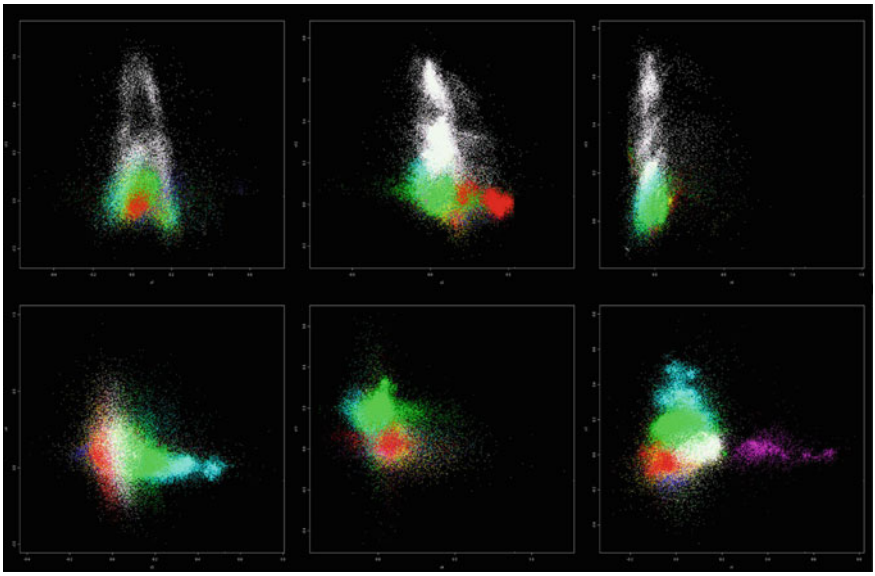


Fig. 16.3 Slices of ES. *Colors* match bacterial phyla. *White*-Proteobacteria, *red*-Actinobacteria, *purple*-Cyanobacteria, *light blue*-Bacteroidetes, *green*-Firmicutes, *blue*-Chloroflexi, *grey*-Acidobacteria, *yellow*-other phyla

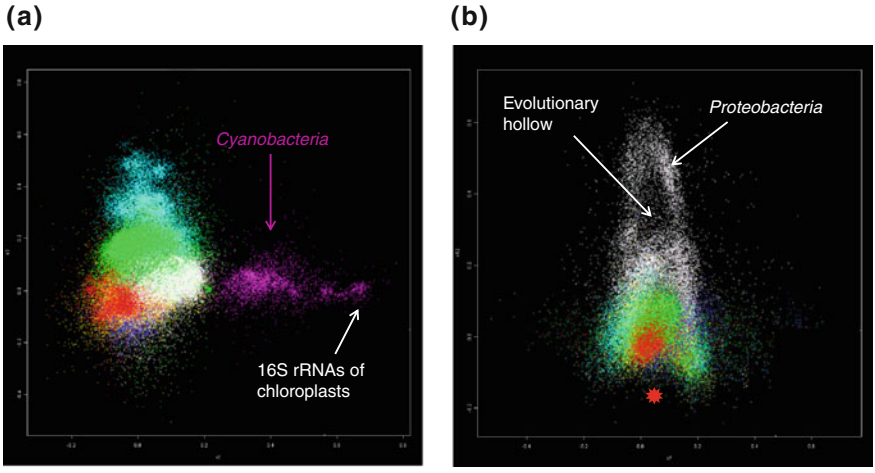


Fig. 16.4 Evolutionary patterns of 16S rRNA gene16S rRNA gene found on ES slices. *Color codes* for bacterial phyla are the same (see Fig. 16.3). The *red asterisk* marks the probable localization of the last common ancestor of bacteria

cellular organelles. In Fig. 16.4a, purple dots mark *Cyanobacteria*. Cyanobacterial "cloud" appears stretched, while its lower apex comes close to the other bacterial phyla. It's quite obvious that the *Cyanobacteria* clearly stand apart from the rest of bacterial phyla. It's worth noting this clear isolation corresponds to some

absolutely unique biological properties of cyanobacteria (e.g., presence of intracellular membranes). As expected, one of its tops faces other bacterial phyla while the opposite one is stacked of the chloroplasts' sequences. This pattern undoubtedly has a clear biological meaning and looks like the "cyanobacterial gun" aimed at the upcoming eukaryotic cell and leading to the formation of its organelles.

16.6 Prospect for Analyzing the Molecular Evolution of Genes

Based on these data we can speculate on how the process of gene evolution can be described in the evolutionary space (Fig. 16.5).

We suggest this process looks quite similar to the Big Bang (expansion of the universe from a singular state). We can mark several regions within the evolutionary space (e.g., areas of ancestral genes; actual rRNA sequences; future modifications of the 16S rRNA that are limited to the functional boundary of the gene). The process of evolution from a common ancestor looks like a gradual filling of the space with evolving variants of the gene. Simultaneously, ancestral gene variants will gradually disappear forming evolutionary hollows. These processes will go on until the gene reaches its functional boundary (the limit after which it loses its functional properties).

An interesting question is what might happen at the functional boundary? Today we can only speculate about it. At this stage the cloud will be "flattened" along the functional boundaries. Since that moment the population of genes undergoes the strong negative selection. It would be quite interesting to calculate the dN/dS statistics (unfortunately this statistics is unavailable for 16S rRNA gene16S rRNA gene) within such regions of space. Further developments may lead to either evolutionary stabilization or "death of the gene" or its saltation. The latter will look like a transition of the gene to a different region of the space (like teleportation), creating a new evolutionary starting point (Fig. 16.5). Latter events can only be seen in evolutionary patterns of the "rapid" genes of "rapid" taxa within "rapid" niches. All aforementioned assumptions are mere hypotheses. Future research may provide more evidence to justify or refute them.

16.7 ES Application in Modern Bacterial Taxonomy

Although the ES concept is under development, we can already talk about the prospects of its application as taxonomic system. Identification of organism will be practically carried out by alignment of corresponding 16S rRNA sequence with reference sequences, forming a simplex. As a result, each organism will be "named", obtaining a set of coordinates in ES. This "naming" of individual 16S rRNAs can be effectively used in a comparative analysis of microbiomes. Microbiomes will be supplied with "taxonomic cards." Each cell within these cards will

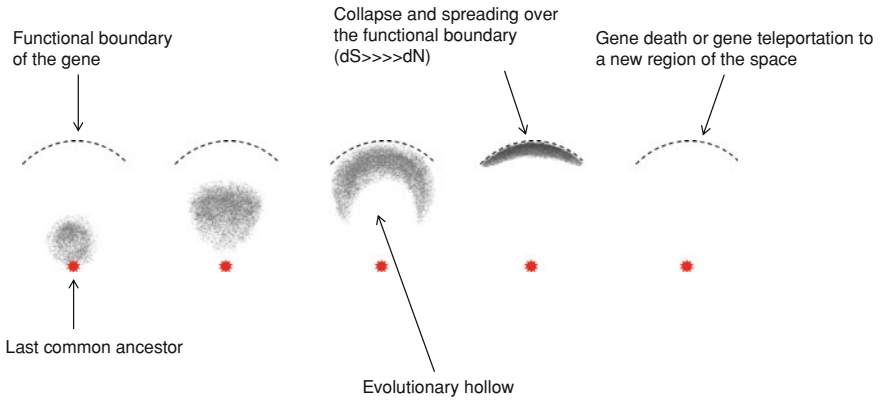


Fig. 16.5 The probable scenario of gene evolution in ES

correspond to the particular 16S rRNA sequence (or group of sequences forming OTU). Practically, these cards can be obtained by unfolding the multidimensional ES into a 2D map (like unfolding the globe to the geographical map). The cell may also have some quantitative characteristics (e.g., cell's color range will indicate the relative abundance of certain organisms in community profile). If we manage to create such a system, it can be applied for quick comparative analysis of microbiomes, using powerful algorithms for graphic patterns recognition.

16.8 Applying ES to Study Succession of Microbiomes

Importantly, further development of ES concept can provide powerful tools for studying adaptive evolution in microbial communities. In our daily practice, we often face the need to study different succession processes in microbial communities exposed to a variety of environmental stresses, including salinization, which is the most powerful environmental factor (Lozupone and Knight 2007). Thus, to demonstrate the potential of ES in ecological studies here we present an example of studying two microbiomes, evolving under the impact of salinization in the field (soil from Kazakhstan salt-marsh) and in short term (150 days) laboratory experiment simulating the conditions of salt-marsh.

Taxonomic structure of soil microbial communities in two experiments is shown on Fig. 16.6a. Clearly, these microbial communities are very diverse, so it is difficult to link particular changes in frequencies of certain microorganisms to the stress factor. Situation is additionally complicated by presence of unidentified sequences (16S rRNAs, which don't have described relatives in databases). These 16S rRNAs can sometimes form a substantial part of the community.

By representing microbial communities in ES we can avoid these problems, treating microbial communities as an integral genetic system and calculating several integral parameters describing their structure and dynamics. Graphical

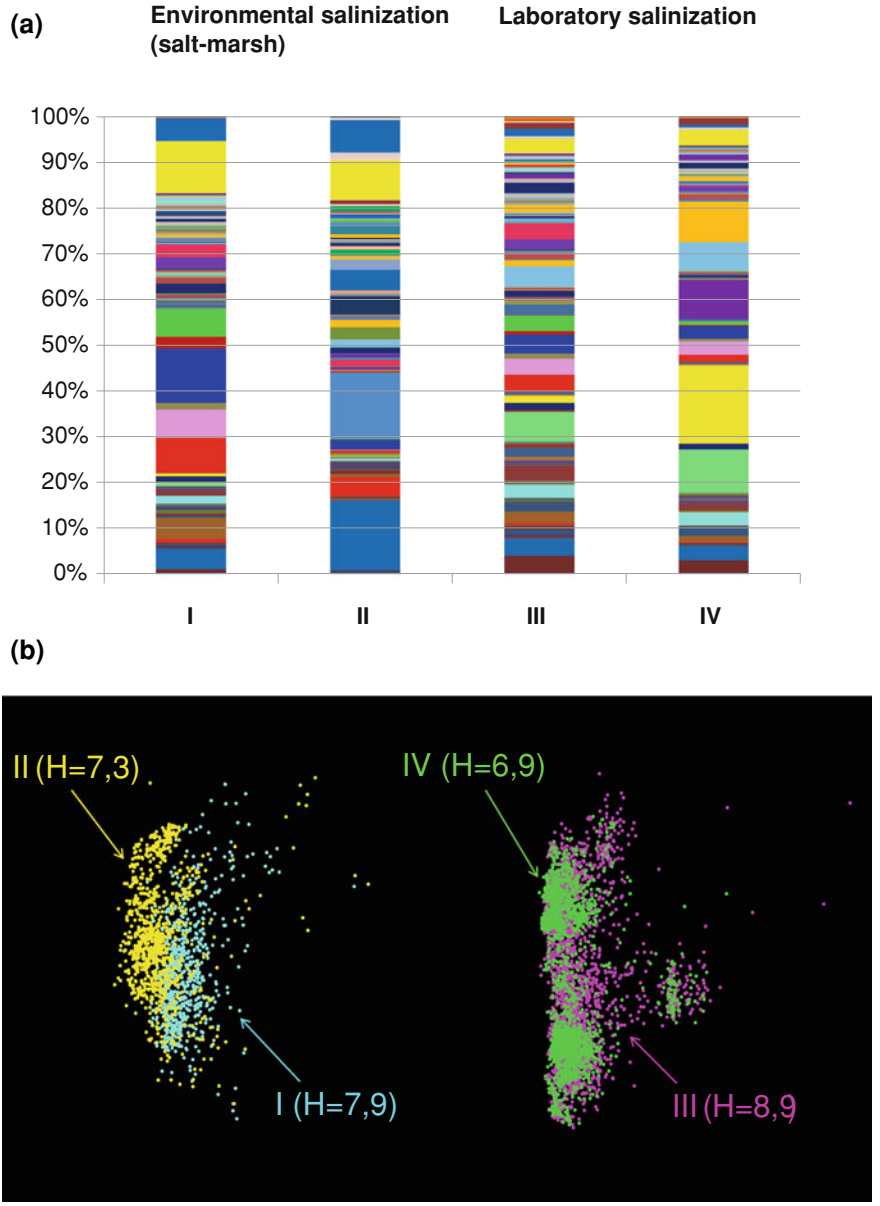


Fig. 16.6 Analysis of microbiomes by use of ES. **a** Taxonomic structure of microbiomes in naturally (*I* non-salinized soil in the vicinity of the salt-marsh, *II* salinized soil from the salt-marsh) and artificially (*III* non-salinized soil in the vicinity of the salt-marsh, *IV* soil *III* salinized in laboratory) salinized environments. Different *colors* mark bacterial families. **b** Representation of the microbiomes in ES (*H* values of Shannon index for microbiomes *I*, *II*, *III*, *IV*)

"portraits" of microbiomes (Fig. 16.6b) give us an opportunity to suggest such integral parameters as shape, volume, density, etc., applying powerful mathematic tools to the analysis of biodiversity. Each geometric parameter will correspond to biological properties of communities. For instance, cloud density will point on the level of genetic heterogeneity in microbial populations.

In this study, we used the simplest integral parameters. We were interested whether there are some similar traits in two processes of salinization. Let's start with microbiome's cloud shapes. We found that after salt stress artificially salted community characterized initially by diffuse arrangements of the dots (meaning high genetic diversity within community), passed in a transitional phase, expressed in cloud collapse, meaning the reduction of diversity in the community and appearance of clearly defined dominant groups (2 areas of point condensation, Fig. 16.6b). In contrast, naturally salted community, due to its long history of adaptations to salinization, was characterized by high levels of diversity, comparable to those exposed in nonsaline soil and clearly expressed "taxonomic shift" (Fig. 16.6b).

As we showed here community shapes may be analogs to biodiversity indexes (e.g., Shannon index, decreasing in transitional artificially salted community and increasing in climax salt-marsh community (The Shannon values are also shown on Fig. 16.6b).

Additionally to structural parameters we can study community dynamics. In some dynamic process each condition (community composition in certain time interval) can be described by central point (the geometric center—the point with average coordinates). It is one of the simplest parameters describing the entire structure of microbiome.

To describe the succession in microbiomes, we can connect two central points (initial and final stages of the community development) with a vector. The length and the direction of this vector will reflect rate and direction of succession. At the same time, the angle between two vectors will point on similarity of the succession traits in different microbiomes (0° for two entirely co-directional patterns of succession and 180° for two opposite patterns). For instance, we can calculate the similarity between natural and artificial processes of salinization. We calculated the angle for our saline samples and it scored at 73°. This indicates certain similarity between two processes of salinization (the angle is not beyond 90° sector), although this similarity is incomplete (the angle differs from zero). Obviously, it cannot be complete because of different development histories of artificial and natural microbiomes.

Surely, for proper interpretation of the angle's value we have to conduct a lot more experiments, using factors with contrast ecological impact.

16.9 Conclusion

In this report we have introduced the first sketch of ES. We call it a sketch as we have obtained low, albeit statistically significant, correlation coefficients ($r = 0.29$, $p < 0.001$) between genetic distances and their ES analogs. Our

preliminary calculations indicate that acceptable correlation will be obtained by increasing the number of dimensions to 40D. To do this, we attempted to expand the simplex using artificially generated 16S rRNA sequences (today we generated nearly 300 "artificial" 16S rRNA sequences extending the basic simplex to 300D). Nevertheless, our first test version clearly demonstrates the ES is already capable as a system for the bacterial taxonomic classification.

We demonstrate the patterns of 16S rRNA gene16S rRNA gene evolution can only be studied in multidimensional spaces. We tried to build such a space using the simplex approach. We found the areas of point accumulation, corresponding to bacterial taxa, as well as empty areas, indicating the probable localization of ancestral sequences.

Evolutionary space has a great potential for application in taxonomic studies, in particular related to the problem of bacterial species (concluding in the lack of universally accepted criteria for typing species for organisms without strict sexual/genetic isolation). In future research, we can speculate on the boarders of the prokaryotic species in the evolutionary space and suggest the novel criteria for their definition in different taxa.

We have also shown that evolutionary space can be used to organize the metagenomic data and to address the main ecological questions about the composition and dynamics of microbiomes, influenced by both environmental and mutational pressures. In future research we can manage to describe the universal evolutionary process undergoing the control of internal factors (mutational process) as well as external factors (natural selection under certain ecological conditions) and thus speculate about the ratio of neutral and adaptive evolutionary events occurring in microbial populations. We hope that ES will be useful to approach this problem.

In this article, we describe only a small part of possible ES applications. Even now we understand that ES has a lot more opportunities in studying bacterial diversity and evolution. But we are still very far from the practical implementation of our ideas. Further development of ES concept requires collaboration of a broad range of specialists beyond the biology itself, including those well versed in complex calculations.

This work was supported by RFBR grants 12-04-00409a and 12-04-01371a.

References

Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl Environ Microbiol 11:3846–3852

Bietz MJ, Lee CP (2009) Collaboration in metagenomics: sequence databases and the organization of scientific work. In: Proceedings of the 11th European conference on computer supported cooperative work, Vienna, 7–11 September 2009

Brandes U, Cornelsen S (2009) Phylogenetic graph models beyond trees. Discrete Appl Math 157:2361–2369

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336

Cardenas E, Cole JR, Tiedje JM, Park J (2009) Microbial community analysis using RDP II (Ribosomal Database Project II): methods, tools and new advances. Environ Eng Res 1:3–9

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141–D145

DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34:W394–W399

Dollo L (1893) Les lois de l'évolution. Bull Soc Belge Geol Pal Hydr 7:164–166

Dolnik AS, Tamazyan GS, Pershina EV, Vyatkina KV, Porozov YuB, Pinaev AG, Andronov EE (2012) The evolutionary space of bacterial 16S rRNA gene v. 1.0. Agric Biol 5:111–120 (in Russian)

Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek 73:25–33

Garrity GM, Lilburn TG (2002) Mapping taxonomic space: an overview of the road map to the second edition of Bergey's manual of systematic bacteriology. WFCC Newsl 35:5–15

Giongo A, Davis-Richardson AG, Crabb DB, Triplett EW (2010) TaxCollector: modifying current 16S rRNA databases for the rapid classification at six taxonomic levels. Diversity 2:1015–1025

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 5:611–618

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 10:4399–4406

Hughes T, Hyun Y, Liberles DA (2004) Visualizing very large phylogenetic trees in three dimensional hyperbolic space. BMC Bioinform 5:48

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267

Kim M, Lee KH, Yoon SW, Kim BS, Chun J, Yi H (2013) Analytical tools and databases for metagenomics in the next-generation sequencing era. Genomics Inform 11:102–113

Kitazoe Y, Kurihara Y, Narita Y, Okuhara Y, Tominaga A, Suzuki T (2011) A new theory of phylogeny inference through construction of multidimensional vector space. Mol Biol Evol 5:812–828

Lee SH, Hwang KS, Lee HR (2006) Embedding operational taxonomic units in three-dimensional space for evolutionary distance relationship in phylogenetic analysis. In: Proceedings of the 5th WSEAS international conference on circuits, systems, electronics, control & signal processing, Dallas, 1–3 November 2006

Lilburn TJ, Garrity GM (2004) Exploring prokaryotic taxonomy. Int J Syst Evol Microbiol 54:7–13

Linnaeus C (1751) Philosophia botanica. Stockholm, Amsterdam. English edition: Linnaeus C (2003) Philosophia Botanica (trans: Freer S). Oxford University Press Inc., New York

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. Proc Natl Acad Sci USA 104:11436–11440

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32:1363–1371

Mantel N, Valand RS (1970) A technique of nonparametric multivariate analysis. Biometrics 26:547–558

Pace NR (2009) Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev 4:565–576

Pommier T, Canbäck B, Lundberg P, Hagström A, Tunlid A (2009) RAMI: a tool for identification and characterization of phylogenetic clusters in microbial communities. Bioinformatics 6:736–742

Pruesse E, Quast C, Knittel K (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 21:7188–7196

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1:283–290

Rudi K, Zimonja M, Næs T (2006) Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of 16S rRNA gene phylogeny. Int J Syst Evol Microbiol 56:1565–1575

Schloss PD, Westcott SL, Ryabin T (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 23:7537–7541

Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. Appl Environ Microbiol 4:1153–1161

The SILVA Ribosomal RNA Database Project (2011) Max Planck Institute for Marine Microbiology, Bremen. http://www.arb-silva.de/no_cache/download/archive/release_104/Exports/. Accessed 24 June 2011

Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE (2012) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J 6:94–103

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Chapter 17
Topopatric Speciation: From Simulations to Theory

David M. Schneider

Abstract In 2009, de Aguiar et al. proposed a neutral model of speciation which successfully predicts empirical patterns of species diversity (de Aguiar et al. 2009). Simulations of the model demonstrate that in absence of natural selection speciation can occur as a consequence of two reinforcing trends: isolation mediated by spacial distance and isolation mediated by genetic incompatibility. Isolation by spacial distance involves a physical mating restriction modeled through a parameter S, representing the search radius for the individuals to find a potential partner. Isolation by genetic incompatibility, on the other hand, is included in the model by preventing mating between individuals whose genetic distance (defined as the number of loci displaying different alleles) exceeds a maximum tolerable difference of G alleles. As speciation in this context does not rely on geographical barriers to be carried out and maintained it could be included in the category of sympatry. However, since mating does not depend only on the genotypes, but individuals also have to be sufficiently close to each other to reproduce, the new term topopatry was adopted to emphasize the role of physical space. In this chapter, we summarize recent advances concerning the construction of a population genetics theory for topopatric speciation. Specifically, we consider first a two-loci model of individuals subjected to mating incompatibilities based on genetic distance, for which a full dynamical description may be provided. Afterward we focus on the influence of mutations, and finally we describe how these outcomes generalize for arbitrarily large genomes.

D. M. Schneider (✉)
Universidade Estadual de Campinas, Rua Seferino Vaz, Campinas–SP, Brazil
e-mail: schneide@ifi.unicamp.br

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_17,

17.1 Different Modes of Speciation

Before the 60's, allopatric speciation dominated the thought of evolutionists as the only mechanism for species formation (Udovic 1980). This paradigm started changing through the contributions of population biologists who, both from theoretical as from experimental angles, started demonstrating that geographical barriers were not necessary for species to be formed and maintained. The process by means of which reproductive isolation could emerge in a single space domain was termed symparty, and the explanation for such counter-intuitive mechanisms involved the concepts of disruptive selection, responsible for the disappearance of intermediate phenotypes, and thus for the blockage of the genetic flux and frequency-dependent selection, which acts stabilizing the population at the marginal phenotypes. Subsequently, new modes of speciation such as parapatric speciation, stasipatry speciation, and marginal sympatry speciation, were described, and all of them referred to an intermediate situation between allopatry and sympatry. This proliferation of terms, in addition to the different opinions on the feasibility of sympatric speciation in nature, created a scenario of controversy, which made it necessary a redefinition of concepts (Greenwood 1984). A new working definition of sympatry was introduced in Kondrashof (1986). In this case, sympatric speciation is not defined in contrast to another, pre-established mechanism (allopatry), but as a process in the course of which the probability of mating between individuals depends on their genotypes only.

In 2009, though, de Aguiar and co-workers proposed a truly different mechanism for speciation (de Aguiar et al. 2009). This mechanism, termed topopatry, does not involve geographical barriers so it has a reminiscence, at least in the antique sense, to symparty. Moreover, natural selection is not present (the process is neutral), but speciation occurs as a consequence of two reinforcing trends: a spacial-mediated reproductive isolation, and a genetic-mediated reproductive isolation. The fact that physical space plays a role in assortativeness, making mating to depend not only on the genotypes, makes the situation strictly different from sympatry and thus deserving a new denomination.

17.2 Topopatric Speciation

Computational simulations of de Aguiar model show that topopatry may be responsible of universal features of biodiversity, such as abundance distributions and species–area relationships (de Aguiar et al. 2009). The key ingredient of the model is the introduction of assortative mating based, not on one as in sympatry, but on two mating restrictions: one in genetic space and the second one in physical space. In the genetic space, an individual can mate only with others whose genome is not too different from its own. The genetic distance is measured by the number of genes bearing different alleles, which in the model are biallelic (the so called

Hamming distance), and the size of the allowed difference is determined by the critical mating genetic distance G. In the physical space, on the other hand, an individual can mate only with others living in a certain neighborhood of its location, in contrast to random (panmictic) mating within the entire population; the size of this neighborhood is determined by the critical mating physical distance S.

Neither of the two restrictions (genetic and physical), when imposed alone, leads to speciation. However, when they are both present, speciation may happen depending on the values of other parameters such as the spatial density of individuals, the mutation rate, the genome size, etc. For a complete description of the parameters involved in the model see (de Aguiar et al. 2009; Baptestini et al. 2013).

But, how do the two restrictions interplay to give rise to speciation? When the critical mating physical distance S is very large, the effect of the genetic mating restriction is to reduce the effective size of the genome, so the average genetic distance is close to G. This situation changes when S is reduced. The net outcome of this constraint is an enhancement of the mutational effect (the effect is similar to the enhancement of mutations in a population with a high percentage of encounters between relatives. Because of positive assortativeness, mutations at different genes tend to be correlated among the individuals, and therefore a given mutation is transmitted with a rate μ instead of μ^2, μ being the mutation rate at that gene). Thus, genes that would be fixed for larger values of S, become now polymorphic, increasing the average genetic distance. When the average genetic distance becomes larger than about $2G$, it is observed that the population can no longer hold itself together and splits in reproductive isolated groups.

A phenomenological description of the pattern observed in the space of critical mating distances (Fig. 17.1), was derived in 2011 by de Aguiar and Bar-Yam. Employing an approach based on the Moran model (1958), these authors obtained for the speciation curve the formula (de Aguiar and Bar-Yam 2011).

$$G_e(S) = \frac{B/2}{1 + exp\left(\frac{\pi^2 N^2 (S - S_{\min})^4}{\gamma^4 \mu^2 B^2 L^4}\right)}, \tag{17.1}$$

where N is the number of individuals, μ the mutation rate, L the size of the physical space available, $S_{\min}$ the minimum radius required for mating, and γ a constant of dimensionality obtained from simulations.

17.3 Population Genetics of Topopatric Speciation

To construct a population genetic theory of topopatric speciation, it is essential to separately understand the effects of each of the ingredients involved, and to evaluate to what extent the mating conditions of haploidism and hermaphroditism restrict the generality of the results. Concerning hermaphroditism, it was already

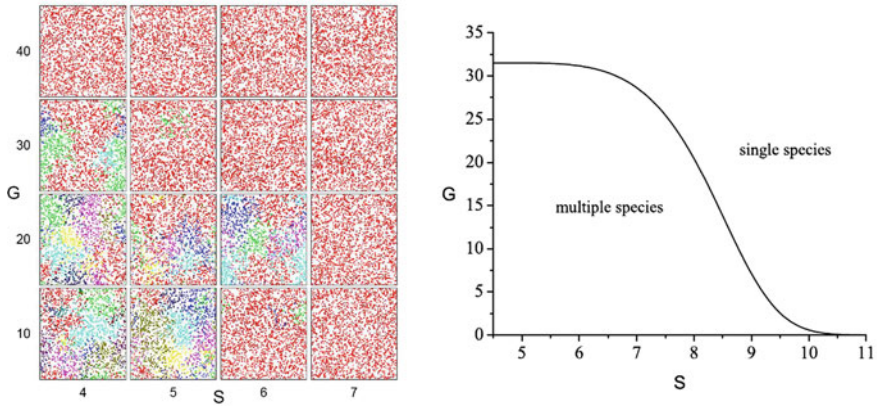


Fig. 17.1 Transition from a single species to multiple species in the space of critical mating distances S and G. In the *left panel*, *colors* refer to different species. *Left* Reprinted from de Aguiar et al. (2009) by permission. *Right* Reprinted from de Aguiar and Bar-Yam (2011) with permission

shown that sex separation do not significantly modify the outcomes of the model (see Baptestini et al. 2013; Schneider et al. 2012). The possible effects of diploidism are under analysis at the moment. In the rest of this work, we discuss one of the central aspects of the topopatric model, namely, the genetic assortative mating. We consider first the dynamics of a population of haploid and hermaphrodite individuals bearing only two biallelic loci, assuming that sexual reproduction is possible between individuals whose haplotypes differ in at most one locus ($G = 1$). In the second place, we describe the effect of mutations on the equilibrium solutions discussed so far, and finally we generalize the results to large genomes with an arbitrary value of G.

17.3.1 Genetic Assortative Mating in Populations of two Bialleic Loci Individuals

Denoting by $\{A, B\}$ and $\{a, b\}$ the alleles at the first and second loci, respectively, the constrain of mating between individuals sharing at least one common allele, leads to the following equations for the evolution of the four haplotype frequencies

$$p_{AB}^{t+1} = \frac{p_{AB}^{t}(1 - p_{ab}^{t})}{1 - 2\Delta^{t}} \tag{17.2}$$

$$p_{Ab}^{t+1} = \frac{p_{Ab}^{t}(1 - p_{aB}^{t})}{1 - 2\Delta^{t}} \tag{17.3}$$

$$p_{aB}^{t+1} = \frac{p_{aB}^{t}(1 - p_{Ab}^{t})}{1 - 2\Delta^{t}} \tag{17.4}$$

$$p_{ab}^{t+1} = \frac{p_{ab}^{t}(1 - p_{AB}^{t})}{1 - 2\Delta^{t}} \tag{17.5}$$

Equations (17.2)–(17.5) display four different types of fixed points, whose characteristics are described in Table 17.1. Only fixed points of types 1 and 2 are stable, representing the real equilibria of the population.

Since $p_{AB} + p_{Ab} + p_{aB} + p_{ab} = 1$, it is possible to give a graphical description of the dynamics by constructing a three-dimensional phase space. We arbitrarily chose the frequencies p_{AB}, $p_{Ab,}$ and p_{aB} as the independent dynamic variables. The constrains $p_{AB} \geq 0$, $p_{Ab} \geq 0$, $p_{aB} \geq 0$, and $p_{AB} + p_{Ab} + p_{aB} \leq 1$ give the phase space the geometry of a tetrahedron having right triangular faces (Fig. 17.2). The top face of the tetrahedron, defined by the equation $p_{AB} + p_{Ab} + p_{aB} = 1$, corresponds to frequencies distributions having $p_{ab} = 0$. $p_{ab} = 1$ implies $p_{AB} + p_{Ab} + p_{aB} = 0$ and is represented by the origin of the coordinate system. Type 1 fixed points are located at four of the six edges of the tetrahedron (colored edges in Fig. 17.2), the points of type 2 are the vertices of the tetrahedron (black circles), type 3 fixed points EU_1 and EU_2 are located at the midpoints of the edges not containing points of type 1 (orange circles) and finally, the center of the tetrahedron houses the type 4 fixed point ES (brown circle).

In contrast to random mating populations, which maintain constant the allele frequencies

$$\tilde{p}_A = p_{AB}^{t} + p_{Ab}^{t} \tag{17.6}$$

$$\tilde{p}_a = p_{aB}^{t} + p_{ab}^{t} = 1 - \tilde{p}_A \tag{17.7}$$

$$\tilde{p}_B = p_{AB}^{t} + p_{aB}^{t} \tag{17.8}$$

$$\tilde{p}_b = p_{Ab}^{t} + p_{ab}^{t} = 1 - \tilde{p}_B, \tag{17.9}$$

Eqs. (17.2)–(17.5) imply that under genetic assortative mating only the combination

$$T = \frac{\tilde{p}_A^{t} - 1/2}{\tilde{p}_B^{t} - 1/2} \tag{17.10}$$

is constant along generations (other combinations of this type are equivalent, by a redefinition of the constant T). For a given value of T, Eq. (17.10) represents a plane in the 3D haplotype space, of equation

Table 17.1 The four types of fixed points of the dynamical system (2–5)

Type	Label	Coordinates	Stability
Type 1. Continuous sets. Two compatible haplotypes have zero frequency; one allele is lost in one locus. The other locus remains polymorphic	E_A	$p_{AB} = p_{Ab} = 0$, $p_{aB} = \lambda_A$, $p_{ab} = 1 - \lambda_A$	Stable
	E_B	$p_{AB} = p_{aB} = 0$, $p_{Ab} = \lambda_B$, $p_{ab} = 1 - \lambda_B$	
	E_a	$p_{aB} = p_{ab} = 0$, $p_{AB} = \lambda_a$, $p_{Ab} = 1 - \lambda_a$	
	E_b	$p_{Ab} = p_{ab} = 0$, $p_{AB} = \lambda_b$, $p_{aB} = 1 - \lambda_b$	
		$\lambda_{A,B,a,b} \in (0, 1)$	
Type 2. Three haplotypes have zero frequency. One allele is lost at both loci	E_{AB}	$p_{AB} = p_{Ab} = p_{aB} = 0$, $p_{ab} = 1$	Stable
	E_{aB}	$p_{AB} = p_{aB} = p_{ab} = 0$, $p_{Ab} = 1$	
	E_{ab}	$p_{Ab} = p_{aB} = p_{ab} = 0$, $p_{AB} = 1$	
	E_{Ab}	$p_{AB} = p_{Ab} = p_{ab} = 0$, $p_{aB} = 1$	
Type 3. Two incompatible haplotypes have zero frequency	EU_1	$p_{Ab} = p_{aB} = 0$, $p_{AB} = p_{ab} = 1/2$	Unstable
	EU_2	$p_{AB} = p_{ab} = 0$, $p_{Ab} = p_{aB} = 1/2$	
Type 4. Equiprobable distribution	ES	$p_{AB} = p_{Ab} = p_{aB} = p_{ab} = 1/4$	Saddle

For equilibria of type 1 and 2, the label subscripts indicate the alleles which are lost

$$p'_{AB} + p'_{Ab} - 1/2 - T(p'_{AB} + p'_{aB} - 1/2) = 0, \tag{17.11}$$

which intersects two of the stable fixed points, being

1. E_B and E_b for $|T| < 1$
2. E_A and E_a for $|T| > 1$
3. E_{AB} and E_{ab} for $T = 1$
4. E_{Ab} and E_{aB} for $T = -1$.

Accordingly, by calculating the value of the constant T from initial conditions, one can determine the plane of motion and thus the equilibrium solution. There is still an ambiguity concerning which of the two intersecting points will be attained. This fact is solved by computing the four allele frequencies and determining the smallest (see Schneider et al. 2014). The allele in the smallest proportion will vanish at equilibrium, so the ambiguity can be eliminated through Table 17.1.

From the biological point of view, taking into account that equilibrium solutions are characterized by the fact that at least one allele vanishes, the results of this assortative mating model can be interpreted as a selection against double polymorphism. In Gavrilets (1999) establishes a second interpretation to models in which assortative mating is based on genetic distance. From that interpretation, it turns out that our results also describe diploid populations subjected to selection against double heterozygotes.

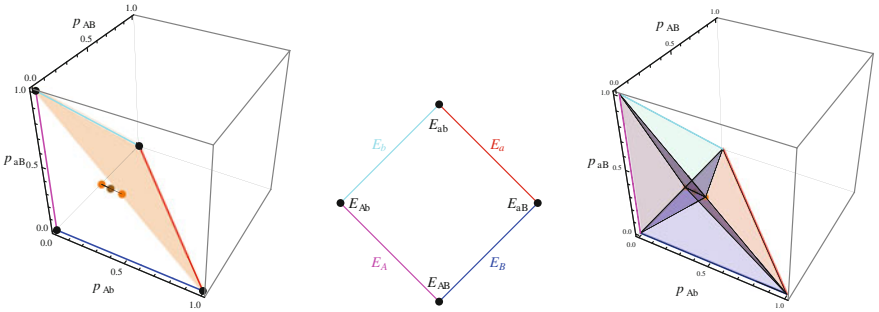


Fig. 17.2 *Left* three-dimensional phase space displaying the four families of equilibrium solutions: E_A (p_{aB} axis, *purple*), E_B (p_{Ab} axis, *blue*), E_a (diagonal on the p_{Ab}–p_{AB} plane, *red*) and E_b (diagonal on the p_{aB}–p_{AB} plane, *cyan*); E_{ab}, E_{Ab}, E_{aB} and E_{AB} (vertices connecting the first family, *black circles*); EU_1 and EU_2 (midpoints of edges of the phase space not containing the first family, *orange circles*); and ES (*center*, *brown circle*). The *shaded light brown* surface represents the top face of the tetrahedral phase space of equation $p_{AB} + p_{Ab} + p_{aB} = 1$. *Middle* Schematic representation of the stable fixed points. *Right* Division of the phase space displaying the basins of attraction of type 1 fixed points

17.3.2 Effect of Mutations on Genetic Assortativeness

When mutations are included in the model, Eqs. (17.2)–(17.5) become

$$p_{AB}^{t+1} = \frac{p_{AB}^t(1-\mu) + p_{ab}^t\mu + \Gamma^t\mu(1-\mu) - p_{AB}^t p_{ab}^t}{1-2\Delta^t} \tag{17.12}$$

$$p_{Ab}^{t+1} = \frac{p_{Ab}^t(1-\mu) + p_{aB}^t\mu - \Gamma^t\mu(1-\mu) - p_{Ab}^t p_{aB}^t}{1-2\Delta^t} \tag{17.13}$$

$$p_{aB}^{t+1} = \frac{p_{aB}^t(1-\mu) + p_{Ab}^t\mu - \Gamma^t\mu(1-\mu) - p_{aB}^t p_{Ab}^t}{1-2\Delta^t} \tag{17.14}$$

$$p_{ab}^{t+1} = \frac{p_{ab}^t(1-\mu) + p_{AB}^t\mu + \Gamma^t\mu(1-\mu) - p_{ab}^t p_{AB}^t}{1-2\Delta^t} \tag{17.15}$$

where μ is the mutation rate at each locus, and $\Gamma^t = p_{Ab}^{2t} + p_{aB}^{2t} - p_{AB}^{2t} - p_{ab}^{2t}$. Equilibrium solutions of Eqs. (17.12)–(17.15) can be obtained analytically. However, the behavior is easier to understand through a graphical construction as shown in Fig. 17.3. Interestingly, the presence of mutations does not affect the planes of constant motion structure of the phase space. Instead, for a given mutation rate the set of equilibria (stable fixed points of type 1 and 2) that for $\mu = 0$ locate at the edges of the phase space, shrink and penetrate the tetrahedrum. For $\mu_{crit} = 1/8$ the stability properties change qualitatively, the equilibrium curves merge in the point corresponding to the equiprobable distribution $p_{AB} = p_{Ab} = p_{aB} = p_{ab} = 1/4$, so the effect of genetic assortative mating disappears and the population behaves as if mating were random.

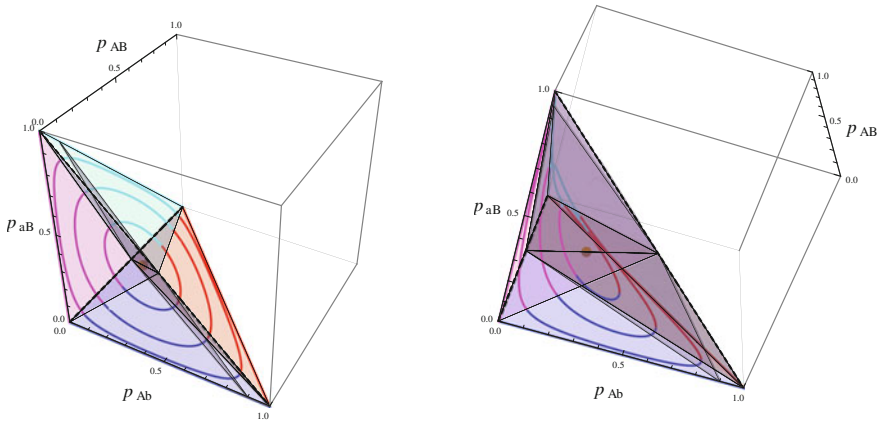


Fig. 17.3 Two visualizations of the equilibrium curves (compare with Fig. 17.2). The curves correspond to different mutation rates, $\mu = 0, 0.01, 0.05, 0.1$ from outside to inside. The *colored regions* represent the basins of attraction of the corresponding equilibrium curves. Plane of constant motion defined by $T = -0.8$ is also shown. In the *right panel*, the angle of observation is modified to highlight the 3D nature of the equilibrium curves

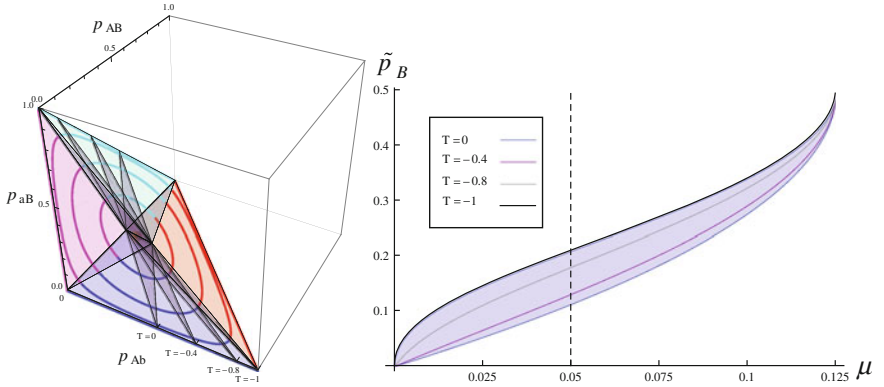


Fig. 17.4 Proportion of allele B at equilibrium in the region of the tetrahedron characterized by having this allele at the minor proportion

An important feature of equilibria for $\mu = 0$ corresponds to the extinction of the initially scarcest allele in the population. When mutations are present, Fig. 17.3 shows that this is not the case. However, is still true that the initially scarcest allele remains being the scarcest allele at equilibrium, in a proportion given by the equilibrium curves. Accordingly, a relevant issue concerns how small is the equilibrium proportion of the initially less abundant allele for a given mutation rate μ. The answer is shown in Fig. 17.4 for the region in which the allele B is in the smallest proportion. In fact, there is not one, but a range of possible $\tilde{p}_B$ equilibrium

values, each one corresponding to a value of T. There is a minimum value at the mid-point of the set E_B ($T = 0$) and a maximum value at the extremes of the sets ($|T| = 1$, points of type 2 E_{AB} and E_{aB}).

17.3.3 Generalization to Large Genomes and Arbitrary Critical Genetic Distance

In order to simplify the notation, let us employ 0 and 1 for the alleles of all genes (the haplotypes *AB*, *Ab*, *Ab,* and *ab* are therefore renamed to 00, 01, 10, and 11, respectively). In the case of B genes the haplotypes assume the form of a string of B digits

$$i = i_1 i_2 \ldots i_k \ldots i_B \quad \text{with } i_k = 0, 1. \tag{17.16}$$

The genetic distance between two haplotypes, i.e., the number of different alleles in the genomes, reads therefore

$$d(i,j) = \sum_{k=1}^{B} |i_k - j_k|. \tag{17.17}$$

There are 2^B different haplotypes, whose frequencies, when only encounters having $d(i, j) \leq G$ are allowed, obey the following evolution equations

$$p_i^{t+1} = \frac{\sum\limits_{\gamma \leq G} p_{i'}^t \, p_{i''}^t \, (1-\mu)^{B-\alpha-\gamma} \mu^{\alpha} \left(\frac{1}{2}\right)^{\gamma}}{1 - \sum\limits_{\gamma > G} p_{i'}^t \, p_{i''}^t}, \tag{17.18}$$

with $\gamma = d(i', i'')$ and $2\alpha = d(i', i) + d(i'', i) - 2d(i', i'')$.

For $\mu = 0$, equilibrium solutions of (17.18) correspond to the fixation of $B - G$ genes (or equivalently, to the disappearance of $B - G$ alleles). The population remains polymorphic at G loci, so individuals can be regarded as having an effective genome of G genes. For $\mu > 0$ and $B \geq 3$, no analytical solutions are available for equilibria. However, it can be demonstrated that for any fixed values of B and G there exists always a critical value of the mutation rate that separates two different behaviors. Below this critical value the behavior is similar to the $\mu = 0$ case, namely, there are $2 \times 2^{B-G}$ different equilibria, each one characterized by the fact that $B - G$ alleles are at a given proportions which are lower with respect to the proportions of the remaining $2G + (B - G)$ alleles. Above $\mu = 0$, instead, the only stable solution is the equiprobable distribution $p_i = 1/2^B$ for every i. The procedure to find this critical value $\mu_{\text{crit}}(B, G)$, involves the linearization of Eq. (17.18) around the equiprobable distribution, and the search of the μ value that makes the second eigenvalue of the stability matrix equal to 1. This reduces to solve the equation

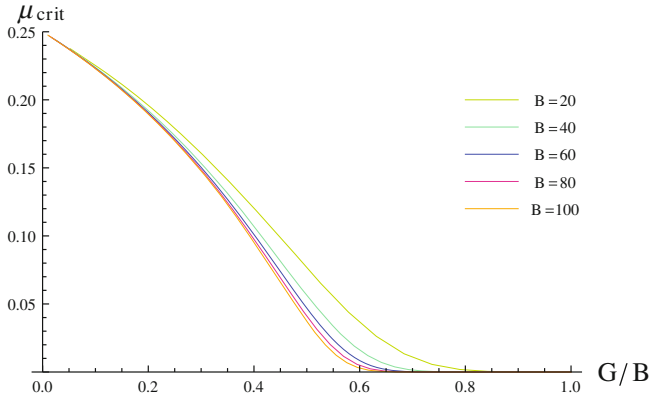


Fig. 17.5 For different genome sizes, μ_{crit} as a function of G/B

$$\sum_{d=0}^{B} c(d,B)A^{B,G}(d,\mu) = 1, \tag{17.19}$$

where $c(d, B)$ are the coefficients satisfying the recursion equation

$$c(k,B) = c(k,B-1) + c(k-1,B-1) \tag{17.20}$$

with $c(0, B) = 1$, $c(1, 1) = -1$ and $c(k, B) = 0$ for $k > B$, and

$$A^{B,G}(d,\mu) = \frac{\mu^d(1-\mu)^{B-d}}{2^{B-1}D}\sum_{k=0}^{d}\binom{d}{k}\left(\frac{1-\mu}{\mu}\right)^k\sum_{i=k}^{G}\binom{B-d}{i-k}\frac{1}{2^i(1-\mu)^i} + \frac{1-D}{2^{B-1}D}, \tag{17.21}$$

with

$$D = 1 - \frac{1}{2^B}\sum_{k>G}\binom{B}{k}. \tag{17.22}$$

Figure 17.5 shows the behavior of μ_{crit} as a function of G/B for different genome sizes. Notice that when B becomes very large, the curves approach to an asymptotic solution $\mu_{\text{crit}}^{\text{asympt}}(G/B)$.

References

Baptestini EM, de Aguiar MAM, Bar-Yam Y (2013) The role of sex separation in neutral speciation. Theor. Ecol. 6:213–223

de Aguiar MAM, Bar-Yam Y (2011) The Moran model as a dynamical process on networks and its implications for neutral speciation. Phys Rev E 84:031901

de Aguiar MAM, Baranger M, Baptestini EM, Kaufman L, Bar-Yam Y (2009) Global patterns of speciation and diversity. Nature 460:384–387

Gavrilets S (1999) A dynamical theory of speciation on holey adaptive landscapes. Am. Nat. 154:1–22

Greenwood PM (1984) What is a species flock? In: Echelle A, Kornfield I (eds) Evolution of fish species rocks. University of Maine at Orono Press, Orono, pp 13–19

Kondrashof AS (1986) Sympatric Speciation: When is it possible?. Biol. J. Linn. Soc. 27:201–233

Moran PAP (1958) Random processes in genetics. Proc. Cam. Phil. Soc. 54:60–71

Schneider DM, do Carmo E, Bar-Yam Y, de Aguiar MAM (2012) Robustness against extinction by stochastic sex determination in small populations. Phys Rev E 86:041104

Schneider DM, do Carmo E, Martins AB, de Aguiar MAM (2014) Toward a theory of topopatric speciation: the role of genetic assortative mating. Physica A 409:35–47

Udovic D (1980) Frequency-dependent selection, disruptive selection, and the evolution of reproductive isolation. Am Nat 106:621–641

Part III
Exobiology and Origin of Life

Chapter 18
A Trip Through Chemical Space: Why Life Has Evolved the Chemistry That It Has

William Bains

Abstract Earth life is built from a quite restricted set of chemicals. Is this an accident of evolution, or are there reasons for the patterns of similarity and diversity in metabolism? I summarize several studies looking at how life has explored "chemical space", and seeking explanations for the nature of biochemistry. The properties of synthetic alternatives to DNA have lead to a hypothesis about what the features that any genetic material must have, providing a rationale for the incorporation of phosphate into DNA. Synthesis and computational studies of possible amino acids have been synergistic in understanding the limits to which the 20 main proteinaceous amino acids are chemically inevitable and the extent to which they are a frozen accident from the origin of life. A broader exploration of chemical space including all of metabolism shows that metabolism is actually crowded into a limited region of the space of possible chemicals. Using a wide dataset of measurement of the toxicity of chemicals, I have shown that this crowding has important implications for how new chemistry can be added to life, which could in principle be developed into a set of constraints on how any metabolism could evolve. Biochemistry is often constrained to specific chemical function, but not always limited to one molecule to carry out that function. These initial results provide hope that there are computationally tractable approaches to understanding why biochemistry is as it is.

18.1 Introduction

Biochemistry is the most complex chemical system known. While geology displays a wider range of elements and atmospheric radical chemistry has at least as complex a web of reactions (see e.g., Hu et al. 2012), only biochemistry has an

W. Bains (✉)
Massachusetts Institute of Technology, 54-1726, 77 Mass Avenue,
Cambridge, MA 02139, USA
e-mail: bains@mit.edu

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2_18,

interlocking network of materials, reactions, and controls that allows it to propagate itself under the direction of a heritable code. To achieve this, life has to have a code, a physical structure, and a complex set of metabolites, including the catalysts needed to make those metabolites.

The dense interconnection of this network of metabolites and their chemistry makes it hard to imagine a different biochemistry. What could take the place of ATP or lysine? Nevertheless, we want to identify what biochemistries are possible, in order both to understand our own origins and to estimate the chances that life could exist elsewhere. Can we educe rules from modern biochemistry that might allow us to rule out such exotica as life using liquid methane as a solvent, or based on silicon as a backbone chemistry (Benner et al. 2004; Committee on the origins and evolution of life 2007; Bains 2004)? Apart from their philosophical implications, such basic questions are starting to have practical importance as research moves toward searching for life on other worlds (Schultz-Mukuch and Irwin 2008) and to building new life in the laboratory (Church and Regis 2012).

In this article, I address the task of seeing if there are approaches that might yield an explanation of why aspects of our biochemistry are as they are, and hence point out aspects that are probably not arbitrary or contingent on an evolutionary path. These are features we could expect to find on life in other worlds, and could inform our understanding of the most primitive form of life and its origin on Earth.

I do not claim here that all aspects of biochemistry have a mechanistic explanation. With Stephen Gould, I agree that if we "rewind the tape of life" a different chemical tune would play, just as different morphological forms would emerge (Gould 1989). But just as the laws of hydrodynamics constrain all large, rapidly-moving aquatic animals to look like big fish, certain basic chemical constraints could drive biochemistry to certain classes of solutions.

This chapter examines several approaches to seeking chemical reasons why the biochemistry we see in terrestrial life is as it is. In the next section, I discuss some general aspects of the chemistry of life. I then summarize work that provides an explanation of why nucleic acids use phosphate backbones, an area of biochemistry that has been illuminated by the systematic attempts to make nucleic acids with other sorts of backbones. Following nucleic acids, I discuss the choice of amino acids used in protein synthesis, which has not only been explored by synthesis of alternatives, but also by computational exploration of what is the smallest number of amino acids that could be used in proteins, and conversely how this "primitive" set could be identified from the universe of possible amino acids. Exploring possible amino acids uses combinatorics as well as chemistry to explore the space of possible biochemistry. Lastly, I discuss my own work on a wider approach to the use of combinatorics to reveal that the metabolic map may be selected as a whole, not as a set of individual pathways, and that some specific features of metabolism, in this case why polyphosphates are universal energy carriers, may be a consequence of the combinatorics of biochemistry.

18.2 Life, Stability, and Solubility

We should start by defining what life is, but there is no robust definition of life and probably one is not possible (Chyba and McDonald 1995; Bedau 2010; Benner 2010; Machery 2012). However, we can argue from our intuitive recognition of living things that life has four core properties (Bains 2013):

1. Structure, that is highly improbable in its environment
2. Dynamic maintenance of that structure, through activity that is characteristic of the organism
3. Occurrence of groups of similar organisms that can be distinguished as a natural group.
4. Substrate-non-determination[1]: living things are determined by an internal code, not (solely) by their external environment.

These all derive from life's use of a code (rather than a pattern) to control, and direct its activities. Thus, we have to find chemistry that explains these features, and provides a mechanism for a code to direct and control its own chemical implementation.

An immediate consequence of the central properties listed earlier is that at least some of the chemicals of life must be chemically stable in a cosmically plausible solvent. This limits biochemistry more than biologists realize. Stability is a reason that silicon cannot be the basis for life on Earth (Bains 2004). Most Si–X bonds are labile to hydrolysis, being easily attacked by water to form Si–OH bonds (an exception of the Si–C bond). Si–H, for example, hydrolyses under human physiological conditions with a half-life of tens of minutes (reviewed in Bains and Tacke 2003). While many metabolites have half-lives much smaller than this, it is clearly impossible to build a genetic material out of chemicals that do not last for one generation. It is this stability issue that meant that arsenic was always a highly improbable backbone component of DNA (Benner et al. 2013).

Solubility is also a key requirement, driven by the observation that life is a dynamic phenomenon. Life has to occur in a dense fluid. A low density fluid, such as a gas, will not be able to solvate the complex polymers needed for catalysis and genetics. A solid does not allow molecules to move and interact with each other. So biochemicals have to dissolve in a dense fluid, which is usually assumed to be a liquid (See Bains 2004 for a short discussion on life in supercritical fluids). Water is the solvent used in terrestrial life, and there is a strong (although not undisputable) argument that it must be the solvent elsewhere. So a minimum stability to aqueous hydrolysis is a basic requirement for potential biochemicals.

It might be thought that these are obvious constraints, but papers continue to be generated suggesting alternative chemistries for life that are completely implausible because of their instability in the stated solvent. We will touch on a more subtle aspect of this in the section later on the amino acids used in proteins.

[1] This was originally termed "substrate independence", but that earlier phrase implied that an organism could grow on anything, which clearly is not true.

18.3 Constraints on the Genetic Material

Life is distinguished from other self-propagating chemical patterns, such as fire, explosions, or crystallization, by its use of a code (Bains 2013). Genes provide a coded set of instructions, not a pattern. A crystal grows because the pattern of ions or molecules on the crystal surface matches that needed to arrange ions or molecules in the solution into a minimum energy configuration. By contrast, the organization of a (terrestrial) lifeform is not a reflection of the organization of its genome, in two senses.

(i) There is no simple mapping of changes in the genome onto changes in the organism. A single base difference in a fly's genome can result in changing a Drosophila from a dipteran with one pair of wings into a mutant with two pairs of wings. Another base difference can change the fly's eyes from black to red. A third can have no apparent effect at all.
(ii) As a consequence, the genome needs a translation system to use its instructions to direct the development of the fly. Differences in the genome are the critical differences between the mutants described earlier, but are only the *cause* of those differences in that sense. They do not *cause* the fly's development in the same sense that a crystal form causes other crystals of the same shape to form from super-saturated solution.

This has important implications for the chemistry of the genetic material, which Benner and others have elaborated (Benner and Hutter 2002; Benner et al. 2004), and whose arguments I will summarize below.

If the genetic material is to code, as distinct from pattern, an organism, then it must have two characteristics:

- It must hold information
- The access and replication of that information must be decoupled from the details of the organism that it encodes.

The first of these is obvious. However, the second is just as important. The code must be uncoupled from the phenotype it describes; its mechanism of replication (DNA synthesis) and decoding (RNA and protein synthesis) must be as unchanging between organisms as is practical. Otherwise a change in the genotype of the cell will inevitably result in a change in the rate of replication or decoding, and the pattern of information in the gene starts to influence the contents of the cell directly rather than through the translation apparatus.

The structure of DNA allows this, because its physicochemical properties are dominated by the sugar-phosphate backbone, not the bases. The bases in DNA can be swapped for nonbiological analogs and, providing the basic spacing of the helix is preserved, DNA and RNA synthesis still allowed (reviewed in Benner et al. 1998; Wojciechowski and Leumann 2011). The order of the bases hardly affects the chemistry of DNA at all, and so the same replication apparatus can, with suitable direction, replicate any DNA regardless of what it codes. By contrast, a wide range of alternatives to the sugar-phosphate backbone have been made, but

only those with charged backbones form regular double helices that recognize the base sequence in other molecules in a predictable way (Benner et al. 1998). Non-phosphate backbones tend to fold in ways that are specific to the sequence of bases in the molecule, more like proteins than DNA. Benner's argument is that such molecules make poor genetic materials precisely because of this sequence sensitivity of their chemistry (Benner and Hutter 2002; Benner et al. 2004).

Does this mean phosphate is an inevitable component of the genetic material? At Earth ambient temperatures the ionic interactions of the DNA backbone, both with itself and with the surrounding solvent, are much stronger than the hydrophobic and hydrogen bonding interactions of the other components of the DNA (Benner and Hutter 2002). Thus, phosphate backbone chemistry (which is the same in all DNA) dominates base chemistry (which changes with sequence). Chemically, DNA is boringly uniform. The need for strong backbone interactions suggests a charged backbone. There are no other mineral ions available that can form stable, charged, covalent bonds: carbonate, silicate, sulfate, sulfite, and borate form uncharged diesters, arsenate esters are too unstable (reviewed in Benner et al. 2013). Among inorganic cations, phosphate seems unique.

However, there are organic polyelectrolytes that could, in principle, replace the sugar-phosphate backbone, with charges originating in carboxylate groups (negative charge) or quaternary amine groups (positive charge). It is the chemical function that is central, not the specific group providing it.

18.4 Constraints on the Proteinaceous Amino Acids

Only 23 amino acids are coded in mRNA (including selenocysteine and pyrrolysine that are coded in a few organisms by modification of stop codon translation (Yuan et al. 2010; Böck et al. 1991), and the N-formyl methionine initiation amino acid used by prokaryotes and their endosymbiotic descendants (Sherman et al. 1985)). However, many other side chains are compatible with protein folding and the translation apparatus, and can add new chemical functionality to proteins. These alternative amino acidsNon-proteinaceous have been successfully incorporated into proteins (Liu and Schultz 2010). In principle, the triplet code could specify 62 amino acids (plus a start and a stop codon). So why does terrestrial life use only the 20 + 3 that are coded, and not others (Weber and Miller 1981; Lu and Freeland 2006)?

The reason(s) are probably different for two classes of amino acids. It plausible that the earliest self-replicating systems used a subset of 10–12 of the proteinaceous amino acids, here termed the Early Amino Acids (the Early group), and that later evolution added to this to expand the chemical functionality of the code with additional amino acids, termed the Late Amino Acids (the Late group) (reviewed in Ilardo and Freeland 2014; Longo et al. 2013).

In general, the Early groups are the amino acids more commonly formed in Miller–Urey type experiments and found in carbonaceous chondrites (Higgs and Pudritz 2009). But not all such easily formed amino acids are used by life. Notable

exceptions are 2-aminobutyrate, 2-aminopentanoate (norvaline), 3-aminopropionate (beta-alanine), and 2,4-dimamobutyrate, which are easily formed but not proteinaceous amino acids. There seems no particular reason for choosing alanine, valine, leucine, and isoleucine and not 2-aminobutyrate or 2-aminopentanoate as small hydrophobic amino acids. Alternatives to lysine are suggested to be more chemically labile to reaction with the peptide backbone (Weber and Miller 1981), but asparagine is labile to deamination and backbone rearrangement reactions (Robinson 2002), and nevertheless is one of the core proteinaceous amino acids. In addition, life makes a range of other amino acids, such as ornithine, citrulline, amino acetoacetate, and homoserine that are nevertheless not incorporated into proteins. So the reason for the selection of the Early Group is not clear.

A more subtle approach is to ask what regions of chemical space are occupied by the proteinaceous amino acids. Chemical space is the systematic description of chemicals in terms not of qualitative structural elements (the classic drawings of chemicals) but quantitative features. Thus, we might plot the amino acids in a 2-dimensional "space" of hydrophobicity and molecular weight, or turn this into a 3-dimensional space by adding a measure of molecular volume (Ilardo and Freeland 2014). Chemical space can have a very large number of dimensions, and as a result be hard to visualize. However, it is mathematically simple to define distances or "volumes" in such high dimensional spaces, whereas comparing lysine with arginine by looking at their structures is fraught with bias, depending on who is looking.

There are many ways to group amino acids in chemical space (Stephenson and Freeland 2013). An initial analysis of the physicochemical properties of the chemical space of actual amino acids suggests that the Early groups are nonrandomly distributed in chemical space (Philip and Freeland 2011). The Late Group expanded the chemical space addressed by amino acids both by extending the chemical space occupied by the Early group and by filling in "gaps" in the Early group's coverage of chemical space (Ilardo and Freeland 2014). Thus, life seems to select specific chemical functions, and evolved new amino acids to provide new function that was not present in the original set. A new origin of life could chose serine and threonine as proteinaceous amino acids providing primary and secondary alcohol function in side chains, respectively, or could chose homoserine and 2-amino-4-hydroxypentanoate instead.

If we expand our search for possible amino acids outside those found in meteorites or Miller–Urey experiments, the number of possible structures becomes very large. For practical reasons we might limit our search to those with (for example) the same number and type of atoms as existing amino acid side chains. Even this generates a vast number of possible structures: Meringer et al. (2013) calculated that there are about 6.5×10^{10} plausible variants of the nine carbons and one nitrogen in tryptophan. Despite this, the search has been started by Freeland and his colleagues (Meringer et al. 2013) using computer methods to generate a large library of possible amino acid structures and then exploring that chemical space, and where the chemistry of life lies within it.

18.5 Self-Consistency and the Molecules of Metabolism

The computational approach to chemical space searches for all the possible chemicals of a particular type. Applying this to amino acids suggests that the proteinaceous amino acids are well spread throughout chemical space, providing an effective spectrum of properties. The exhaustive computational search of chemical space can be generalized to looking at the complete chemical space of all molecules. The advantage of such approaches is that they can address all types of biochemistry, including the proposed "RNA world" that preceded our current protein-based biochemistry (Lazcano and Miller 1996; Orgel 1998), and even entirely alien chemistry. The disadvantage is that chemical space is infinite, and even if we limit ourselves to small molecules still contains millions or billions of compounds.

Despite the computational task, we have been pursuing a general combinatorial approach to characterizing the chemical space of actual biochemicals, using a very similar approach to (Meringer et al. 2013). The software used is different, but generates broadly similar numbers of molecules (Bains and Seager 2012, 2013). As in (Ilardo and Freeland 2014), we have concentrated on simple measures of molecular structure that can be captured as a number, rather than qualitative descriptions. This allows the distribution of biochemicals to be compared to the distribution of all possible chemicals in the dimensions being considered.

One interesting parameter is the so-called Redox Ratio (*Rr* (Bains and Seager 2012, 2013)). This is a measure of the saturation of the atoms in a molecule, and is very easily calculated from molecular structure or formulae. In short,

$$Rr = 1 - \frac{\sum S_A}{\sum S_H}$$

where S_A is the number of hydrogen atoms in the molecule and S_H is the maximum number of hydrogen atoms that could be bonded to the nonhydrogen atoms in that molecule. This is a measure of the redox state of the molecule, cast in structural terms rather than the more conventional energetic or electrochemical terms (Bains and Seager 2012). *Rr* is independent of the size and composition of the molecule, and is independent of whether the atoms can have charges or not. *Rr* for methane is 0, for methanol is 1/3, and for carbon dioxide is 1. In the case of methanol, *Rr* is calculated as follows. The number of hydrogens in methanol is 4. The molecule contains one carbon atom, which can be joined to a maximum of four hydrogens, and one oxygen, which can be joined to a maximum of two hydrogen atoms. Thus, $\sum S_H = 6$. Thus, $Rr = 1 - 4/6 = 2/6 = 1/3$.

The definition of *Rr* only makes sense for molecules (or their salts) built of elements that form stable covalent bonds to hydrogen, and for which forming a bond to hydrogen is reduction. This is roughly equivalent to elements with an Allred and Rochow electronegativity ≥ 1.8 (Allred and Rochow 1958). These are B, Ga, C, Si, Ge, N, P, As, Sb, O, S, Se, Te, F, Cl, Br, and I. While the *Rr* of a

series of similar compounds correlates roughly with their standard electrode potential (as illustrated in Bains and Seager 2012), the comparison is not really valid, as standard electrode potential relates to how compounds *react*, whereas *Rr* relates to their *structure*. Thus, a compound can have an *Rr* on its own, but its electrode potential relates to its *reduction*.

We introduced *Rr* as a simple, pragmatic description of the oxidation state of a molecule based on its bonding, as contrasted to its energetics. The redox ratio *Rr* is not meant to be profound, but convenient for such structural comparison. However, it brings up an important feature of the chemistry of life.

18.5.1 Rr and Chemical Space

If we plot the distribution of the number of chemicals in *Rr*-defined chemical space, we see that molecules are overwhelmingly at *Rr* values between 0.45 and 0.8 (Fig. 18.1). This is equivalent to saying that almost all discrete, covalent molecules are of intermediate redox state, and have some hydrogen atoms in. This is hardly a revelation to chemists, but is a problem for life on a small planetary body such as Earth, Venus, or Titan. Carbon is likely to be present as its most oxidized form (CO_2) or most reduced form (CH_4) on a dense body with low surface pressure and moderate surface temperature. This is because CO_2 or CH_4 are the thermodynamically most stable forms of carbon, and in a dense environment where reactions can happen fast, carbon atoms will end up in their most thermodynamically stable form. Whether environmental carbon is present as CO_2 (or as carbonate), or as CH_4 depends on the redox state of the environment. In oxidizing environments such as the surface of Earth, Venus, and Mars, CO_2 is the more stable, as illustrated in Fig. 18.2 (note that an "oxidized" environment is one where the majority of compounds are in an oxidized state: this is distinct from an "oxidizing" environment, such as the modern day surface of the Earth, where elemental oxygen is present as a powerful oxidizing agent). In reducing environments such as the surface of Titan, CH_4 is the most likely carbon compound.

Life needs a lot of different chemicals to build metabolism, catalysts, and the genetic material. In order to acquire this chemical diversity, life needs to access the chemical space in which there are a lot of chemicals. This means the region of chemical space where the chemicals are intermediate *Rr*. Therefore, of necessity, life must take compounds of $Rr = 0$ (CH_4) or $Rr = 1$ (CO_2) and convert them into compounds of intermediate redox state ($0 < Rr < 1$).

This is the reason that photosynthesis on Earth produces oxygen. To generate biomass (summarized here as a generic carbohydrate structure CH_2O), life must reduce the carbon in CO_2 and hence must oxidize something else:

$$CO_2 + H_2O + X \rightarrow CH_2O + 2XO$$

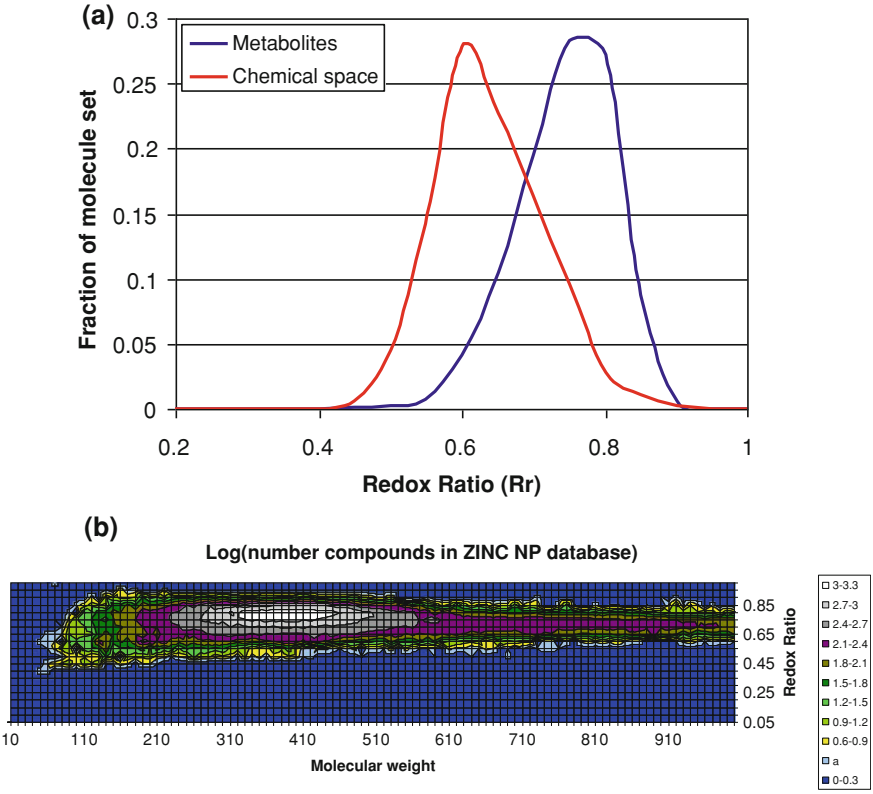


Fig. 18.1 *Rr* chemical and biochemical space. **a** Distribution of redox ratio (Rr) in biochemicals and in the chemical space of possible biochemicals. X axis: Rr. Y axis—fraction of all molecules in a set having that Rr. The data sets: "Chemical Space" is the set of chemicals from which biochemicals are selected, generated as described in (Bains and Seager 2012): in short, atoms and bonds are combined in all combinatorial possibilities compatible with valencies of the atoms involved and with some basic rules to bias towards compounds that are not reactive, are stable in water, and are made of elements likely to be in biochemicals, here C, H, O, N, S as S(II) and P as P(V) bonded to at least two oxygens. Here all molecules of 3–9 nonhydrogen atoms are used, a total of 2,000,584 molecules. "Metabolites" is the set of all natural products downloaded from the ZINC database (https://zinc.docking.org/browse/catalogs/naturalproducts), with the exception of the Tinji data set. **b** Distribution of metabolites as a function of molecular weight (X axis) and redox ratio (Y axis). Contour map of Log_{10}(number of molecules) in bins of 10 Dalton mass units and 0.05 *Rr* units. Molecules from ZINC database as described in part A. Note that the display is truncated at M.wt. = 1,000 Daltons for display. (**b**) shows that the distribution in (**a**) is not a feature of any particular molecular weight class of metabolites

The *Chloroplastida* have developed chemistry whereby X is the oxygen in water, so the reaction becomes

$$CO_2 + H_2O \rightarrow CH_2O + O_2$$

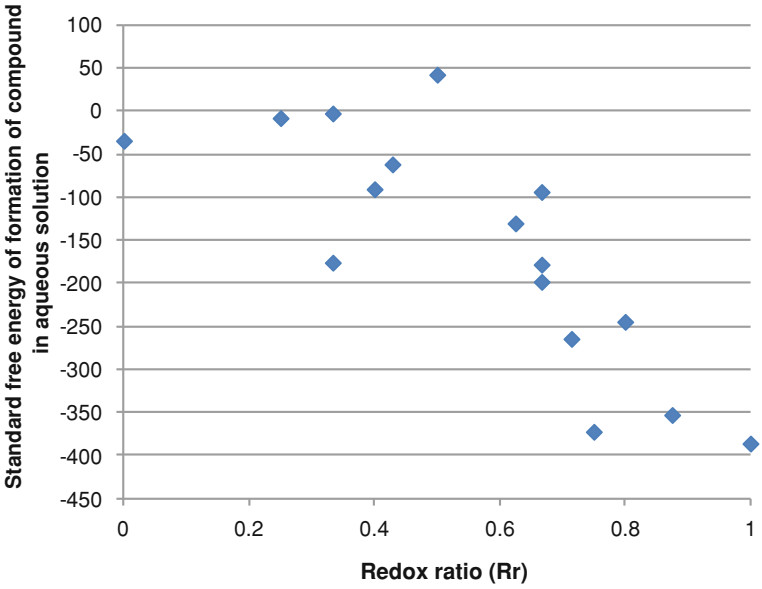


Fig. 18.2 Free energy versus redox ratio. Plot of redox ratio (x axis) versus free energy of formation per carbon atom (Y axis), calculated for 1, 2 and 3 carbon compounds containing only carbon, hydrogen, and oxygen in aqueous solution. Data from (Amend and Shock 2001; Chase 1998). Compounds (in order of increasing *Rr*) are methane, ethane, methanol, propane, ethanol, propanol, ethylene, propanoic acid, lactic acid, formaldehyde, acetic acid, glycolic acid, formic acid, malonic acid, oxalic acid, carbon dioxide

The reaction above requires substantial energy input, which is provided by the abundant energy source of sunlight. Given that abundant source of energy, making molecular oxygen is the simplest way of getting rid of the "surplus" oxygen atoms in water and carbon dioxide so that life can make biomass with an intermediate *Rr*.

Life is parsimonious, and we might expect photosynthesising organisms to use as little energy as practical to reduce CO_2 to biomass. There is a hint that this is so. The distribution of the *Rr* of actual molecules in biochemistry is biased toward high *Rr* (oxidized) compounds, as shown in Fig. 18.1. Under the oxidized conditions of the surface of the Earth, this is an energy-conserving strategy, as compounds with high *Rr* require less energy to synthesize from CO_2 than highly reduced compounds with low *Rr* (Fig. 18.2). Life seems to have preferentially occupied the volume of chemical space near to its source of carbon—CO_2—a strategy that minimizes the amount of energy needed to take inorganic carbon and turn it into simple organic molecules. Whether this is the reason for the offset in the curves in Fig. 18.1, or whether there are other reason(s), are not shown by this data.

Restrictions on *Rr* are a very weak constraint on the chemistry of life. However, *Rr* may be used to probe more subtle but more powerful limitations, related to the inter-connectivity of the chemistry of a cell.

18.5.2 Rr and Toxicity

Biochemical pathways and their protein catalysts do not exist in isolation. The participants in the chemistry of the cell are immersed in the milieu containing the other participants.[2] The evolution of metabolism by adding new pathways to a basic core requires that all the new metabolites be consistent with the existing metabolism, an increasing number of constraints as metabolism is expanded to include (for example) the Late group amino acids (Ilardo and Freeland 2014; Longo et al. 2013). One would expect the difficulty of adding a new chemical to metabolism to be related to how similar the chemical is to ones already there (illustrated in simple cartoon form in Fig. 18.3). This might also impose patterns on metabolism similar to the ones found when searching for patterns of amino acids: new metabolites would have to be similar to existing metabolites (because they are derived from them), but not too similar (to avoid cross-talk with existing metabolic pathways and their enzymes).

To explore the effect of adding a "new" chemical on biochemistry, I looked at the effect of nonnatural chemicals on living systems. The overwhelming effect of synthetic chemicals on living systems is to interfere with their operation, i.e., to poison them. Sometimes, this is because of specific pharmacological effect (i.e., a high-affinity blockade of a particular mechanism, which is how drugs work), sometimes it is the result of chemical reactivity. However, many chemicals are mildly toxic for no particularly obvious reason. It is this low level of toxicity that would be expected from a chemical that, by chance, interfered with a range of metabolic processes through its similarity to biochemicals. Because of the crowding of biochemicals into the highest *Rr* corner of chemical space, we might expect this effect to be particularly pronounced for high *Rr* chemicals.

I tested this with several datasets of the toxicity of a wide range of molecules on some whole, aquatic organisms commonly used to test chemicals for toxicity (Wang 1990; Cronin et al. 2004; Das and Roy 2014; Schultz 1999; Schultz et al. 2005). Aquatic organisms were chosen to avoid issues of how the chemicals were delivered. Whole organism toxicity was chosen as the endpoint because the point is to probe the integrated effect of a chemical on all aspects of biological function, not on one specific pathway or mechanism. Within this requirement, species were chosen for which there were measurements of the toxic effects of at least 50 chemicals, from diverse chemical structural families, available in the public literature, that were measured in a comparable fashion. Thus, for example, organisms used almost exclusively to measure one type of compound (such as chlorinated hydrocarbons or detergents) were not considered, and organisms in which many compounds were studied, but in different or poorly defined ways were also not suitable. I do not claim the resulting choice is comprehensive, but the

[2] In a prokaryotic cell internal compartment, anyway. In compartmentalized eukaryotic cells, each compartment only 'sees' a subset of metabolism.

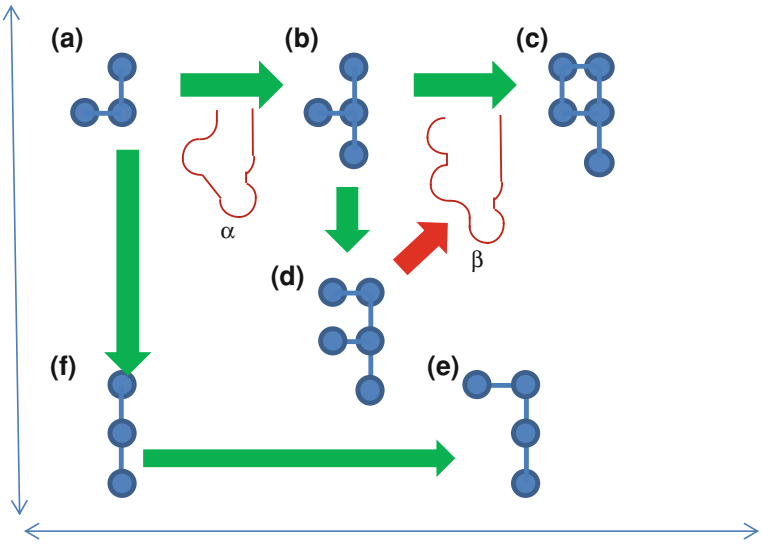


Fig. 18.3 Cartoon of argument for chemical space selection. Cartoons of six metabolites in a 2-dimensional chemical space. Horizontal and vertical axes represent two dimensions of a "chemical space". Metabolite *a* is converted to metabolite *b* through the action of a specific catalyst α, and thence to metabolite *c* through the action of β. *a*, *b*. and *c* are close to each other in a 2D chemical space. The organism can benefit from a new metabolite, *d* or *e*. *d* can be made from *b* in one metabolic step, but is so close to *b* and *c* that it is likely to interfere with their synthesis or function (here illustrated by competitive inhibition of the synthesis of *c*). There is selective advantage to evolve the longer pathway through metabolite *f* to metabolite *e* to avoid conflict with existing chemistry

computational methods are sufficiently simple that they can readily be applied to other data sets. The organisms used and toxicity data sources are listed in Table 18.1.

Three sample results are shown in Fig. 18.4. There is a clear negative correlation of toxicity with *Rr* for the unicellular organism *Chlorella,* a weak but plausible negative for the multicellular plant *Lemna*, and no correlation for the multicellular animal *Gammarus*. A negative correlation of *Rr* with EC_{50} means that a high *Rr* is associated with a low EC_{50}, i.e., with a compound that has its effect at a low concentration, i.e., compounds that are more potent toxins. So the expected correlation is observed for *Chlorella* and *Lemna*, but not for *Gammarus*.

Rather than burden the reader with many graphs of the sort shown in Fig. 18.4, the results of 8 such correlations are summarized in Table 18.2. In some, but not all cases, there is a negative correlation between *Rr* and toxicity. To summarize, there is a highly statistically significant correlation between *Rr* and global toxicity for *Chlorella*, *Scenedesmus*, *Tetrahymena*, Yeast, and *Lemna,* but no significant correlation for the two invertebrates *Gammarus* and *Pteronarcys*. Also included in Table 18.2 is the correlation between the Redox Ratio and toxicity for a set of herbicides acting against *Lemna*. There is only marginal significance in this

Table 18.1 Data sources for toxicity data

Data set	Organism	Toxicity endpoint	Number of data points (chemicals)	Source	Comment
Chlorella	*Chlorella vulgaris*	Death	91	(Cronin et al. 2004)	
Lemna—nonherbicides	*Lemna gibba* and *Lemna minor*	Lack of growth/leaflet reduction	149	US EPA IPM database at http://www.ipmcenters.org/Ecotox/DataAccess.cfm, and (Larson et al. 2008; Blackman et al. 1955; Wang 1990; Cowgill et al. 1991; Tong and Hongjun 1997; Hanson and Solomon 2004; Brain et al. 2004a, b; Sharma et al. 1997; Pillard and DuFresne 1999; Ramirez Toro et al. 1988; Kirby and Sheahan 1994; van de Plassche et al. 1999; Boudreau et al. 2003; Qi et al. 2011; McConkey et al. 1997; Xu et al. 1988; Berends et al. 1999; Caux et al. 1988)	Toxicity derived from 7-day and 14-day frond (leaflet) number tests
Lemna—herbicides			174	US EPA IPM database at http://www.ipmcenters.org/Ecotox/DataAccess.cfm	
Tetrahymena	*Tetrahymena pyriformis*	Death	334	(Schultz 1999; Schultz et al. 2002, 2005; Akers et al. 1999)	
Scenedesmus	*Scenedesmus obiquus*	Growth inhibition/death (cell numbers)	62	(Yan et al. 2005; Saçan et al. 2007; Wang et al. 2008, Lu et al. 2001; Ma et al. 2002, 2006, 2007; Tadros et al. 1994; Zhang et al. 2012; Ma 2005; Kulacki and Lamberti 2008; Geoffroy et al. 2002; Li et al. 2005)	Dataset heavy on chlorinated and nitrated aromatic compounds
Yeast	*Saccharomyces cereviseae*	Growth inhibition	197	http://dtp.nci.nih.gov/yacds/download.html	Mostly drug-like molecules: data points selected as per legend

(continued)

Table 18.1 (continued)

Data set	Organism	Toxicity endpoint	Number of data points (chemicals)	Source	Comment
Gammanarus	Combined data from *G. fasciatus*, *G. lacustris* and *G. Pseudolimnaeus*	Death	132	*Acquatic acute toxicity data downloaded from* http://www.cerc.usgs.gov/data/acute/acute.html	96 h toxicity
Pteronarcys	*Pteronarcys californica*	Death	52	http://www.cerc.usgs.gov/data/acute/acute.html	96 h toxicity

Sources of data for this study. Data was filtered to collect toxicity endpoints that were, as far as practical, the same for different studies. Only studies that provided a quantitative half-effect concentration were included. All EC_{50} values were as reported in the relevant papers or databases except those for *S. cereviseae*, where EC_{50} values were calculated from the raw inhibition data downloaded from http://dtp.nci.nih.gov/yacds/download.html. This very large data set was filtered to exclude organometallic compounds, mixtures or salts other than halogen or alkali metal salts (so as to exclude the possibility that toxicity effects were due to the counter ion rather than the test compound), compounds for which growth inhibition at the highest concentration was <50 % or for which the growth inhibition at the lowest concentration was >50 %, and compounds for which the range of calculated EC_{50}s across the 13 strains tested in this data set (calculated as [maximum(EC_{50})-minimum(EC_{50})]/average(EC_{50}) was >1. The resulting data set represented well defined organic compounds with EC_{50}s within the experimentally measured concentrations and consistent toxicity across a range of *S. cereviseae* strains

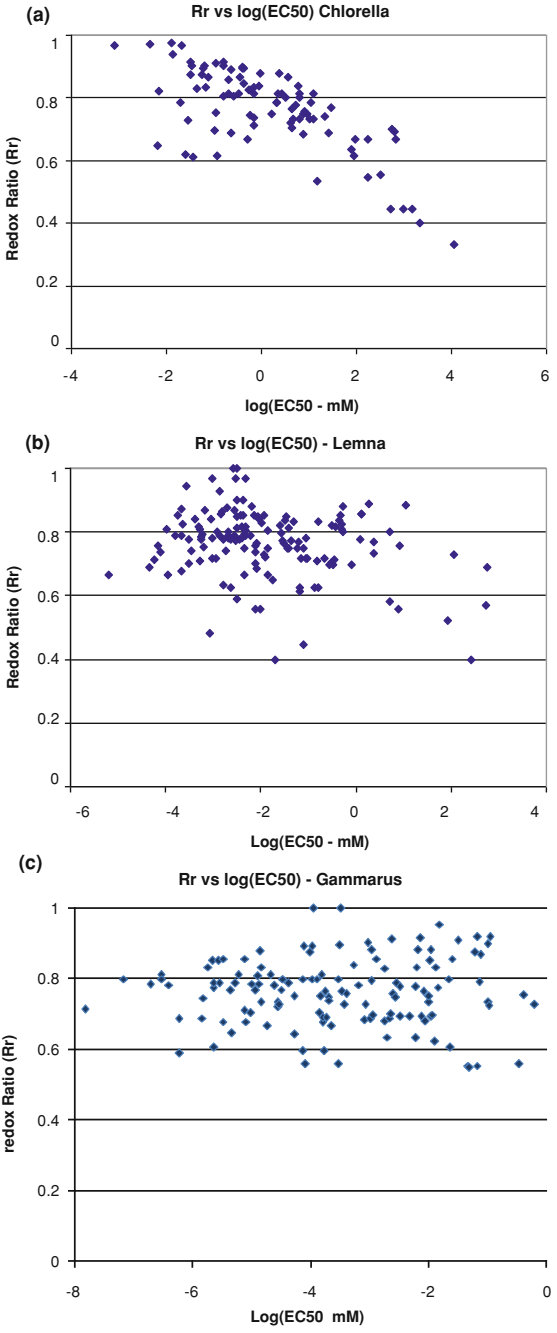


Fig. 18.4 Plot of *Rr* versus toxicity in three example organisms. Plot of log of the measured EC_{50} (concentration required to produce half-effect in toxicity assay, usually growth cessation) versus redox ration (*Rr*). Only three species are shown, **a** *Chlorella* (which shows a strong correlation between *Rr* and toxicity) and **b** *Lemna* (which shows a weak correlation) and **c** *Gammarus* (which shows no correlation)

Table 18.2 Correlation statistics

		Chlorella	Tetrahymena	Scenedesmus	Yeast	Lemna (nonhebicides)	Lemna (herbicides)	Pteronarcys	Gammarus
	N	91	334	62	197	174	149	52	132
	Rr	–0.710	−0.576	−0.384	−0.2313	−0.254	−0.182	−0.0309	−0.0025
	P(*Rr*)	3.65E-17	6.5e-31	2.06E-03	1.07E-3	7.1E-4	0.0259	0.827	0.976
	M.wt.	−0.787	−0.641	0.132	−0.0294	−0.338	−0.474	−0.4366	−0.4848
	P(M.wt.)	2.2E-20	4.1E-40	0.0303	0.0681	4.79E-06	1.00E-09	1.21E-03	3.77E-09
Multiple regression	Multiple regression	0.857	0.698	0.396	0.234	0.414	0.474	0.4536	0.5046
	P(MR)	4.42E-26	1.24E-48	0.00604	4.2E-3	1.09E-9	3.59E-10	0.0035	5.87E-09
	Constant (C)	5.451 (0.484)	2.574 (0.215)	1.830 (1.626)	−0.621 (0.402)	1.695 (0.835)	−0.834 (0.902)	−5.796 (1.737)	−2.082 (1.011)
	M.Wt factor (μ)	−0.0109 (0.0012)	−0.00868 (0.00087)	0.00127 (0.00160)	0.00012 (0.0002)	−0.00305 (0.00070)	−0.0082 (0.0013)	−0.0101 (0.0028)	−0.0115 (0.0017)
	Rr factor (ρ)	−4.510 (0.729)	−2.397 (0.343)	−6.096 (1.937)	−1.333 (0.4023)	−3.357 (1.054)	0.1696 (1.274)	2.2666 (2.2348)	2.466 (1.339)

Correlation statistics for *Rr* and molecular weight (M.wt) versus toxicity EC_{50}. Top line—species. N = number of data points for that data set. *Rr* = Pearsons Correlation Coefficient ("correlation") of gross toxicity with Redox Ratio for that data set. P(*Rr*) = probability of obtaining that correlation by chance. M.wt = correlation of molecular weight with gross toxicity for that data set. P(M.wt) = probability of obtaining that correlation by chance. Multiple regression = correlation of both *Rr* and M.wt in linear multiple regression with toxicity. Note that this is always positive. P(MR) = ANOVA probability of obtaining that correlation by chance. C, μ and ρ are the factors for the multiple regression function, of the form

$$Tp = C + \mu \cdot \mathrm{M} \cdot \mathrm{Wt} + \rho \cdot Rr$$

where Tp is our prediction of the log EC_{50} value, and C, μ and ρ are constants. Values in brackets are the standard errors of the values of these constants

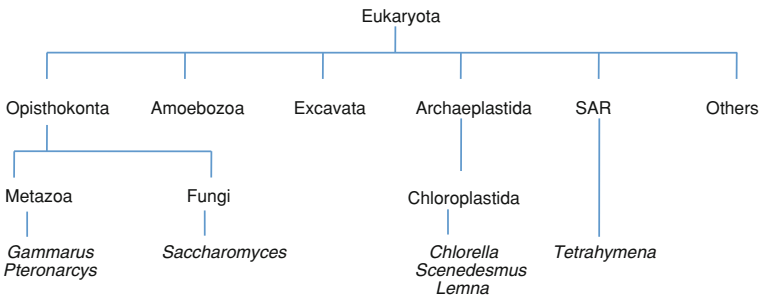


Fig. 18.5 Phylogeny of species studied. Top level of classification of eukaryotes, flagging the classifications of the species used in this study. After (Adl et al. 2012)

correlation—given the number of correlations reported in this chapter, a significance level of $p = 0.01$ should be a minimum threshold for even considering a correlation the product of anything other than chance.

The fact that such a simple measure of molecular structure is correlated with toxicity is surprising. The pattern of correlation between species is also surprising. Despite fungi being more closely related to animals than to plants (Fig. 18.5), *Saccharomyces* and the three unicellular plant species show modest correlation while the two animals do not. Despite *Tetrahymena* being an out-group to this collection of organisms, it shows the same correlation of *Rr* with toxicity as the two unicellular plants.

Lemna, *Gammarus* and *Pteronarcys*, the most weakly correlating species, are multicellular, whereas the other four species are unicellular. This suggests that multicellularity may be a factor in determining correlation between *Rr* and toxicity. This is not implausible, as multicellular organisms have macroscopic differentiation between body parts specifically to defend themselves from environmental chemicals. It is possible therefore that a correlation of *Rr* with toxicity is being masked by the multicellular organism's defenses against toxins, which would be directed against the metabolites made by other organisms, and so would selectively defend against high *Rr* compounds because these are most likely to be metabolites.

18.5.3 Molecular Weight and Toxicity

To provide an initial probe into whether multicellularity might be a factor in the different correlation of *Rr* with toxicity, I have also correlated toxicity with molecular weight. Molecular weight is generally associated with an increase in molecular "stickiness"—the tendency of molecules to interact with protein targets in a nonspecific way (reviewed in the context of drug screening in Rawlins 2010; Leeson and Springthorpe 2007; Hopkins et al. 2006). It might therefore also be

correlated with broad toxicity (i.e., mechanism-independent) endpoints. However, molecular weight is not a good way to distinguish biological molecules from other molecules (at least below 500 daltons), and so would not be the basis for a multicellular organisms's defenses against toxins.

Table 18.2 shows that there is indeed a negative correlation between EC_{50} and molecular weight, and it is seen in both unicellular and multicellular eukaryotes studied here. The exception is *Scenedesmus*, which has a positive correlation of E_{50} with molecular weight (i.e., bigger molecules have higher ED_{50}, i.e., are less toxic). However, the number of compounds tested in *Scenedesmus* is quite small, and the correlation is consequently of marginal significance. A larger data set might provide more robust evidence for or against this anomaly.

The correlation of both Molecular Weight and Redox Ratio with toxicity suggests that a combination of the two might be more predictive than either alone. Table 18.2 probes this, and finds that generally it is not: although *Rr* and M.Wt. are not strongly correlated (Correl coefficient = + 0.2181 for a set of 5117 compounds, including those used here), the correlation of *Rr* and M.wt combined with toxicity is only marginally better than the correlation of one or other parameter with toxicity. The complete lack of additivity between *Rr* and molecular weight in predicting toxicity suggests that the correlation of *Rr* with toxicity and molecular weight with toxicity may arise from related mechanisms.

18.5.4 Rr, Molecular Weight and Toxicity: So What?

The observations above are not just a rather inefficient way to detect toxicity, but hint at global relationships between the components of metabolism. The cartoon in Fig. 18.3 explain the argument. Any new chemical entity that we wish to add to metabolism has to have a minimum difference from *all* the components of metabolism in order to be *easily* accepted into that metabolism. That is not to say that two almost identical molecules *cannot* coexist in the same metabolism, but it requires especially adaptation of all the proteins that interact with both metabolites to make that happen. Because chemical space at high *Rr* is more crowded with biochemicals, this means it is harder to add a new molecule to metabolism if that molecule has a high *Rr*. The uniform spacing of the amino acids found by Freeland's work (Philip and Freeland 2011; Stephenson and Freeland 2013; Meringer et al. 2013) could therefore not only be a result of selection *for* wide exploration of chemical space, but also selection *against* the "new" amino acids being too similar. The same selection would apply to any other metabolite. One might therefore be able to predict ab initio how diverse a metabolism must be to function, an ambitious project that would nevertheless shed light on what a "minimal replicator" would have to achieve to sustain life.

18.6 Why ATP?

Lastly, the combinatorial approach to exploring chemical space can have some unexpectedly specific implications, illustrated by a possible explanation of why life chose nucleoside triphosphates as the universal energy currency (Bains 2013).

Twenty five years ago, Westheimer laid out the argument why polyphosphates had to be the energy carrier, and not other common anionic species (Westheimer 1987). Phosphate esters are unusually stable to hydrolysis (Wolfenden 2006), and the negative charge on phosphate esters and diesters, which provided the structural stability to DNA described earlier, also provides a thermodynamic drive to their hydrolysis. So phosphates provide a form of trapped free energy, their hydrolysis strongly thermodynamically favored but kinetically inhibited until released in an enzyme reaction.

However, phosphate esters are not the only metabolites that can be hydrolysed to release energy, and in any case—why *ATP* rather than (say) the much simpler phospho-enol-pyruvate (PEP), which packs and even greater energetic punch on hydrolysis? Figure 18.6 shows the free energy of hydrolysis of a range of metabolites, and those widely used in energy-transferring reactions all cluster around not only a specific chemical type, but also a specific energy.

There may be a clue as to the reason for the choice of this energy, and hence of ATP, in the combinatorics of selecting biochemistry from chemical space (Bains 2013). The $\sim$600 molecules of primary metabolism are around 10–11 atoms in size (not counting hydrogens), or are made of modules containing 10 or 11 atoms. If we take 10 atoms as a typical size of a "basic metabolite", there are $\sim 3.5 \times 10^7$ possible 10-atom molecules that can be made of C, N, O, S(II) and P(V) (Bains and Seager 2012), from which life has selected $\sim$600, a selection of 1 in 1.7×10^5. This is an estimate of the probability that we see the same group of 600 chemicals in a wide range of living organisms, rather than either a random mixture of all the $\sim 3.5 \times 10^7$ possible chemicals, or different random subsets of 600 chemicals in each organism, *assuming* (and this is unrealistic) that the molecules are equally likely. If we equate this information with the Boltzmann entropy needed to select 1 molecule from a pool of 1.7×10^5, the entropy is –91 J/mole, and hence requires a free energy of $\sim$27 kJ/mol at 298 K (see Hitchcock and Lovelock 1967) for more details on Boltzmann entropy). The free energy of hydrolysis of Mg.ATP is 25–27 kJ/mol under intracellular conditions, very similar to the energy needed to select the chemicals of life from a random space of all possible chemicals (Bains 2013). One might hypothesize that nucleoside triphosphates were selected as a universal energy carrier because their energy of hydrolysis was best suited to the need to select a small number of chemicals from the large chemical space of possible chemicals. If this is the explanation (and I emphasize this is very speculative), then if life had taken a different metabolic turn, Acetyl-CoA could have been the universal energy carrier, but PEP could not.

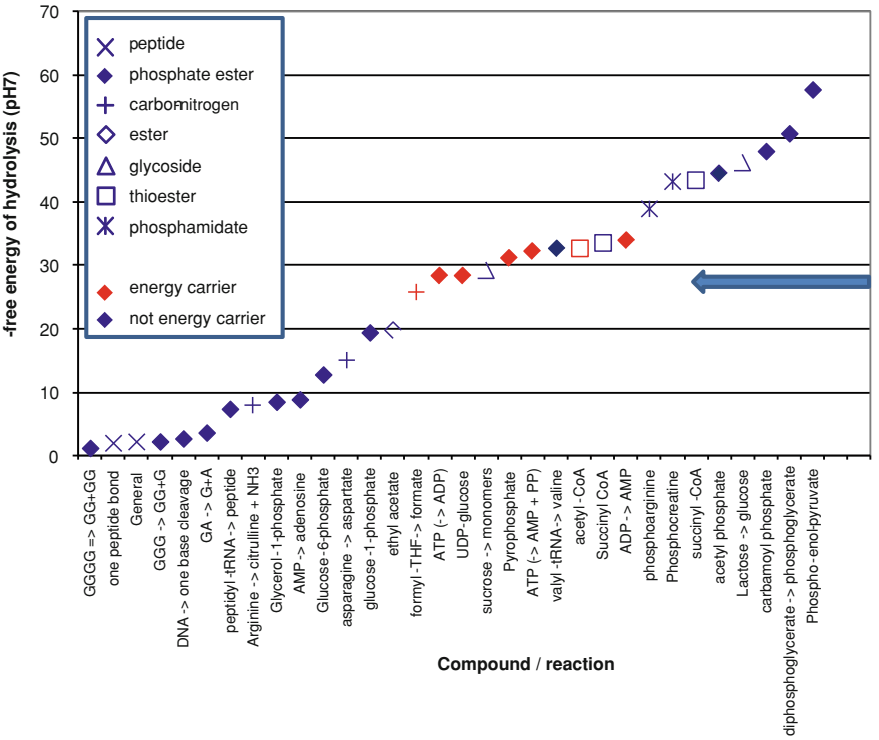


Fig. 18.6 Free energy of hydrolysis of some metabolites. Free energy of hydrolysis of some metabolites, categorized by the class of bond cleaved and by whether they serve as a general energy carrier in primary metabolism or not. An energy carrier is defined as a metabolite that provides substantial input free energy into a wide range of metabolic reactions and processes, e.g., tetrahydroformate as a donor of methyl groups in a number of C_1-transferring reactions. *Horizontal arrow*—Boltzmann entropic energy for selection of metabolism, based on calculation described in the text. Data from (Martin and Russell 2007; Metzler and Metzler 2001)

18.7 Conclusion

Our trip through chemical space has not answered the question in the title, about why life has evolved the chemistry that it has. However, the examples have shown that such a question is becoming answerable. The twin tools of synthetic chemistry to make variants of the chemistry of life, and computational chemistry to explore the chemical space in which biochemistry swims, are starting to reveal clues about what aspects of life are historically contingent accidents and what may be the result of selection. In particular, I believe that treating biochemistry as an integrated whole, rather than a set of disconnected reactions and pathways, will prove to be critical to unraveling how life arose and the path it took to get to the biosphere we see around us.

Acknowledgements My thanks to Janusz Petkowski (ETH, Zurich) for many useful discussions, to the attendees of the Gordon Research Conference on Synthetic Biology (Summer 2013) for helpful comments, to Sara Seager (MIT) for the continued and unstinting support, to Pierre Pontarotti and the organizers of the 17th Evolutionary Biology Meeting (2013, Marseilles) for inviting me, and to Alan Wilson (Llhasa Ltd) for not believing a word of it.

References

Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, le Gall L, Lynn DH, McManus H, Mitchell EAD, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Schoch CL, Smirnov A, Spiegel FW (2012) The revised classification of eukaryotes. J Eukaryot Microbiol 59:429–514

Akers KS, Sinks GD, Schultz TW (1999) Structure–toxicity relationships for selected halogenated aliphatic chemicals. Environ Toxicol Pharmacol 7:33–39

Allred AL, Rochow EG (1958) A scale of electronegativity based on electrostatic force. J Inorg Nucl Chem 5:264–268

Amend JP, Shock EL (2001) Energetics of overall metabolic reactions of thermophilic and hyperthermophilic archaea and bacteria. FEMS Microbiol Ecol 25:175–243

Bains W (2004) Many chemistries could be used to build living systems. Astrobiology 4:137–167

Bains W (2013) What do we think life is? A simple illustration and its consequences. Int J Astrobiol (in press)

Bains W, Seager S (2012) A combinatorial approach to biochemical space: description and application to the redox distribution of metabolism. Astrobiology 12:271–281

Bains W, Seager S (2013) Correction: a combinatorial approach to biochemical space: description and application to the redox distribution of metabolism. Astrobiology 13:792

Bains W, Tacke R (2003) Silicon chemistry as a novel source of chemical diversity in drug design. Curr Opin Drug Discov Devel 6:526–543

Bedau MA (2010) An Aristotelian account of minimal chemical life. Astrobiology 10:1011–1020

Benner S, Bains W, Seager S (2013) Models and standards of proof in cross-disciplinary science: the case of arsenic DNA. Astrobiology 13:510–513

Benner S, Battersby TR, Eschgfäller B, Hutter D, Kodra JT, Lutz S, Arslan T, Bäschlin DK, Blättler M, Egli M, Hammer C, Held HA, von Krosigk U, Lutz MJ, Macpherson LJ, Moroney SE, Müller E, Nambiar KP, Piccirilli JA, Switzer C, Vögel JJ, Richert C, Roughton AL, Schmidt J, Schneider KC, Stackhouse J (1998) Redesigning nucleic acids. Pure Appl Chem 70:263–266

Benner S, Hutter D (2002) Phosphates, DNA, and the search for nonterrean life: a second generation model for genetic molecules. Bioorg Chem 30:62–80

Benner S, Ricardo A, Carrigan M (2004) Is there a common chemical model for life in the universe? Curr Opin Chem Biol 8:672–689

Benner SA (2010) Defining life. Astrobiology 10:1021–1030

Berends AG, Boutonnet JC, Rooij CGD, Thompson RS (1999) Toxicity of trifluoroacetate to aquatic organisms. Environ Toxicol Chem 18:1053–1059

Blackman GE, Parke MH, Garton G (1955) The physiological activity of substituted phenols. I. relationships between chemical structure and physiological activity. Arch Biochem Biophys 54:45–54

Böck A, Forchhammer K, Heider J, Leinfelder W, Sawers G, Veprek B, Zinoni F (1991) Selenocysteine: the 21st amino acid. Mol Microbiol 5:515–520

Boudreau TM, Sibley PK, Mabury SA, Muir DGC, Solomon KR (2003) Laboratory evaluation of the toxicity of perfluorooctane sulfonate (PFOS) on *Selenastrum capricornutum*, *Chlorella*

vulgaris, *Lemna gibba*, *Daphnia magna*, and *Daphnia pulicaria*. Arch Environ Contam Toxicol 44:0307–0313
Brain RA, Johnson DJ, Richards SM, Hanson ML, Sanderson H, Lam MW, Young C, Mabury SA, Sibley PK, Solomon KR (2004a) Microcosm evaluation of the effects of an eight pharmaceutical mixture to the aquatic macrophytes *Lemna gibba* and *Myriophyllum sibiricum*. Aquat Toxicol 70:23–40
Brain RA, Johnson DJ, Richards SM, Sanderson H, Sibley PK, Solomon KR (2004b) Effects of 25 pharmaceutical compounds to *Lemna gibba* using a seven-day static-renewal test. Environ Toxicol Chem 23:371–382
Caux PY, Weinberger P, Carlisle DB (1988) A physiological study of the effects of triton surfactants on *Lemna minor* L. Environ Toxicol Chem 7:671–676
Chase MWJ (1998) NIST-JANAF thermochemical tables, fourth edition. J Chem Phys Ref Data Monograph Number 9:1–1951
Church GM, Regis E (2012) Regenesis: how syntehtic biology will reinvent nature and ourselves. Basic Books, New York
Chyba CF, McDonald GD (1995) The origin of life in the solar system: current issues. Ann Rev Earth Planet Sci 23:215–249
Committee on the origins and evolution of life (2007) The limits of organic life in planetary systems. National Research Council, Washington
Cowgill UM, Milazzo DP, Landenberger BD (1991) The sensitivity of *Lemna gibba* G-3 and four clones of *Lemna minor* to eight common chemicals using a 7-day test. Res J Water Pollut Fed 63:991–998
Cronin MTD, Netzeva TI, Dearden JC, Edwards R, Worgan ADP (2004) Assessment and modeling of the toxicity of organicchemicals to *Chlorella vulgaris*: development of a novel database. Chem Res Tox 17:545–554
Das RN, Roy K (2014) Predictive modeling studies for the ecotoxicity of ionic liquids towards the green algae *Scenedesmus vacuolatus*. Chemosphere 104:170–176
Geoffroy L, Teisseire H, Couderchet M, Vernet G (2002) Effect of oxyfluorfen and diuron alone and in mixture on antioxidative enzymes of *Scenedesmus obliquus*. Pestic Biochem Physiol 72:178–185
Gould SJ (1989) Wonderful life: the burgess shale and the nature of history. Norton & Co, New York
Hanson ML, Solomon KR (2004) Haloacetic acids in the aquatic environment part I: macrophyte toxicity. Environ Pollut 130:371–383
Higgs PG, Pudritz RE (2009) A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology 9:483–490
Hitchcock DR, Lovelock JE (1967) Life detection by atmospheric analysis. Icarus 7:149–150
Hopkins AL, Mason JS, Overington JP (2006) Can we rationally design promiscuous drugs? Curr Opin Struct Biol 16:127–136
Hu R, Seager S, Bains W (2012) Photochemistry in terrestrial exoplanet atmospheres. i. photochemistry model and benchmark cases. Astrophys J 761:166
Ilardo MA, Freeland S (2014) Testing for adaptive signatures of amino acid alphabet evolution using chemistry space. J Syst Chem 5: doi:10.1186/1759-2208-5-1
Kirby MF, Sheahan DA (1994) Effects of atrazine, isoproturon, and mecoprop on the macrophyte *Lemna minor* and the alga *Scenedesmus subspicatus*. Bull Environ Contam Toxicol 53:120–126
Kulacki KJ, Lamberti GA (2008) Toxicity of imidazolium ionic liquids to freshwater algae. Green Chem 10:104–110
Larson JH, Frost PC, Lamberti GA (2008) Variable toxicity of ionic liquid-forming chemicals to *Lemna minor* and the influence of dissolved organic matter. Environ Toxicol Chem 27:676–681
Lazcano A, Miller SL (1996) The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. Cell 85:793–798

Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discovery 6:881–890
Li X, Ping X, Xiumei S, Zhenbin W, Liqiang X (2005) Toxicity of cypermethrin on growth, pigments, and superoxide dismutase of *Scenedesmus obliquus*. Ecotoxicol Environ Saf 60:188–192
Liu CC, Schultz PG (2010) Adding new chemistries to the genetic code. Ann Rev Biochem 79:413–444
Longo LM, Lee J, Blaber M (2013) Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. Proc Natl Acad Sci 110:2135–2139
Lu G-H, Yuan X, Zhao Y-H (2001) QSAR study on the toxicity of substituted benzenes to the algae (*Scenedesmus obliquus*). Chemosphere 44:437–440
Lu Y, Freeland S (2006) On the evolution of the standard amino-acid alphabet. Genome Biol 7:102
Ma J (2005) Differential sensitivity of three cyanobacterial and five green algal species to organotins and pyrethroids pesticides. Sci Total Environ 341:109–117
Ma J, Lu N, Qin W, Xu R, Wang Y, Chen X (2006) Differential responses of eight cyanobacterial and green algal species, to carbamate insecticides. Ecotoxicol Environ Saf 63:268–274
Ma J, Wang P, Chen J, Sun Y, Che J (2007) Differential response of green algal species *Pseudokirchneriella subcapitata*, *Scenedesmus quadricauda, Scenedesmus obliquus*, *Chlorella vulgaris* and *Chlorella pyrenoidosa* to six pesticides. Polish J Environ Stud 16:847–851
Ma J, Zheng R, Xu L, Wang S (2002) Differential sensitivity of two green algae, *Scenedesmus obliqnus* and *Chlorella pyrenoidosa*, to 12 pesticides. Ecotoxicol Environ Saf 52:57–61
Machery E (2012) Why i stopped worrying about the definition of life… and why you should as well. Synthese 185:145–164
Martin W, Russell MJ (2007) On the origin of biochemistry at an alkaline hydrothermal vent. Philos Trans R Soc B Biol Sci 362:1887–1926
McConkey BJ, Duxbury CL, Dixon DG, Greenberg BM (1997) Toxicity of a pah photooxidation product to the bacteria photobacterium phosphoreum and the duckweed *Lemna gibba*: effects of phenanthrene and its primary photoproduct, phenanthrenequinone. Environ Toxicol Chem 16:892–899
Meringer M, Cleaves HJI, Freeland SJ (2013) Beyond terrestrial biology: charting the chemical universe of α-amino acid structures. J Chem Inf Model 53:2851–2862
Metzler DE, Metzler CM (2001) Biochemistry—the chemical reactions of living cells, vol 1, 2nd edn. Harcourt Academic Press, San Diego
Orgel LE (1998) The origin of life—a review of facts and speculations. Trends Biochem Sci 23:491–495
Philip GK, Freeland S (2011) Did evolution select a nonrandom “alphabet” of amino acids? Astrobiology 11:235–240
Pillard DA, Dufresne DL (1999) Toxicity of formulated glycol deicers and ethylene and propylene glycol to *Lactuca sativa*, *Lolium perenne*, *Selenastrum capricornutum*, and *Lemna minor*. Arch Environ Contam Toxicol 37:29–35
Qi P, Wang Y, Mu J, Wang J (2011) Aquatic predicted no-effect-concentration derivation for perfluorooctane sulfonic acid. Environ Toxicol Chem 30:836–842
Ramirez TORO GI, Leather GR, Einhellig FA (1988) Effects of three phenolic compounds on *Lemna gibba* G3. J Chem Ecol 14:845–853
Rawlins P (2010) Current trends in label-free technologies. Drug Discov World 2010:17–26
Robinson NE (2002) Protein deamidation. Proc Nat Acad Sci USA 99:5283–5288
Saçan MT, Özkul M, Erdem SS (2007) QSPR analysis of the toxicity of aromatic compounds to the algae (*Scenedesmus obliquus*). Chemosphere 68:695–702
Schultz-Mukuch D, Irwin LN (2008) Life in the universe: expectations and constraints, 2nd edn. Springer, Berlin
Schultz TW (1999) Structure—toxicity relationships for benzenes evaluated with *Tetrahymena pyriformis*. Chem Res Toxicol 12:1262–1267

Schultz TW, Cronin MTD, Netzeva TI, Aptula AO (2002) Structure—toxicity relationships for aliphatic chemicals evaluated with *Tetrahymena pyriformis*. Chem Res Toxicol 15:1602–1609

Schultz TW, Netzeva TI, Roberts DW, Cronin MTD (2005) Structure—toxicity relationships for the effects to *Tetrahymena pyriformis* of aliphatic, carbonyl-containing, α, β-unsaturated chemicals. Chem Res Toxicol 18:330–341

Sharma HA, Barber JT, Ensley HE, Polito MA (1997) A comparison of the toxicity and metabolism of phenol and chlorinated phenols by *Lemna gibba*, with special reference to 2,4,5-trichlorophenol. Environ Toxicol Chem 16:346–350

Sherman F, Stewart JW, Tsunasawa S (1985) Methionine or not methionine at the beginning of a protein. BioEssays 3:27–31

Stephenson J, Freeland S (2013) Unearthing the root of amino acid similarity. J Mol Evol 77:159–169

Tadros MG, Philips J, Patel H, Pandiripally V (1994) Differential response of green algal species to solvents. Bull Environ Contam Toxicol 52:333–337

Tong Z, Hongjun J (1997) Use of duckweed (*Lemna minor* L.) growth inhibition test to evaluate the toxicity of acrylonitrile, sulphocyanic sodium and acetonitrile in China. Environ Pollut 98:143–147

van de Plassche EJ, de Bruijn JHM, Stephenson RR, Marshall SJ, Feijtel TCJ, Belanger SE (1999) Predicted no-effect concentrations and risk characterization of four surfactants: Linear alkyl benzene sulfonate, alcohol ethoxylates, alcohol ethoxylated sulfates, and soap. Environ Toxicol Chem 18:2653–2663

Wang C, Lu G, Tang Z, Guo X (2008) Quantitative structure-activity relationships for joint toxicity of substituted phenols and anilines to *Scenedesmus obliquus*. J Environ Sci 20:115–119

Wang W (1990) Literature review on duckweed toxicity testing. Environ Res 52:7–22

Weber A, Miller S (1981) Reasons for the occurrence of the twenty coded protein amino acids. J Mol Evol 17:273–284

Westheimer FH (1987) Why nature chose phosphates. Science 235:1173–1178

Wojciechowski F, Leumann CJ (2011) Alternative DNA base-pairs: from efforts to expand the genetic code to potential material applications. Chem Soc Rev 40:5669–5679

Wolfenden R (2006) Degrees of difficulty of water-consuming reactions in the absence of enzymes. Chem Rev 106:3379–3396

Xu Y, Lay JP, Korte F (1988) Fate and effects of xanthates in laboratory freshwater systems. Bull Environ Contam Toxicol 41:683–689

Yan X-F, Xiao H-M, Gong X-D, Ju X-H (2005) Quantitative structure–activity relationships of nitroaromatics toxicity to the algae (*Scenedesmus obliguus*). Chemosphere 59:467–471

Yuan J, O'donoghue P, Ambrogelly A, Gundllapalli S, Sherrer RL, Palioura S, Simonović M, Söll D (2010) Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. FEBS Lett 584:342–349

Zhang W, Zhang M, Lin K, Sun W, Xiong B, Guo M, Cui X, Fu R (2012) Eco-toxicological effect of carbamazepine on *Scenedesmus obliquus* and *Chlorella pyrenoidosa*. Environ Toxicol Pharmacol 33:344–352

Index

P. Pontarotti (ed.), *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*, DOI: 10.1007/978-3-319-07623-2,