

Pierre Pontarotti *Editor*

---

# Evolutionary Biology—A Transdisciplinary Approach

 Springer

# Evolutionary Biology—A Transdisciplinary Approach

Pierre Pontarotti  
Editor

# Evolutionary Biology—A Transdisciplinary Approach

 Springer

*Editor*

Pierre Pontarotti 

IHU Marseille MEPHI

Aix Marseille Univ IRD, APHM

Marseille, France

SNC5039 CNRS

Paris, France

ISBN 978-3-030-57245-7

ISBN 978-3-030-57246-4 (eBook)

<https://doi.org/10.1007/978-3-030-57246-4>

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

For the 13th time now, we publish a book on concepts and applications in evolutionary biology.

Our aim is to capture the evolution and progress of this field, and the Evolutionary Biology Meeting in Marseilles is a perfect basis to do so. 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.

The book chapters have been selected from the meeting presentations and from propositions born by the interaction of meeting participants.

The readers of the evolutionary biology books as well as the meeting participants would perhaps like to see a shift in the evolutionary biology concepts, which they have witnessed year after year in the different meetings and also in the books. However, the fact that the chapters of the books are selected from a meeting enables the quick diffusion of the novelties.

We would like to emphasise that the 13 books are complementary to each other and should be considered as tomes.

Marseille, France  
June 2020

Marie Hélène Rome  
A.E.E.B. Director

Pierre Pontarotti  
A.E.E.B. and CNRS

**Acknowledgements** We would like to thank all the authors and the reviewers of the different chapters.

We thank the sponsors of the meeting. Aix Marseille Université, CNRS, ECCOREV Federation, Conseil Départemental 13, ITMO, Ville de Marseille.

We wish to thank the A.E.E.B. team for the organisation of the meeting.

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

# Contents

<b>1</b>	<b>Genetic and Morphological Differentiation of Common Toads in the Alps and the Apennines</b> . . . . .	<b>1</b>
	Jan W. Arntzen, Wouter de Vries, Daniele Canestrelli, and Iñigo Martínez-Solano	
<b>2</b>	<b>Molecular Phenotypes as Key Intermediates in Mapping Genotypes to Fitness</b> . . . . .	<b>15</b>
	Aditya Ballal, Constantin D. Malliaris, and Alexandre V. Morozov	
<b>3</b>	<b>A Practical Guide to Orthology Resources</b> . . . . .	<b>41</b>
	Paul de Boissier and Bianca H. Habermann	
<b>4</b>	<b>Protein Recoding Through RNA Editing: Detection, Function, Evolution</b> . . . . .	<b>79</b>
	Eli Eisenberg	
<b>5</b>	<b>Most Successful Mammals in the Making: A Review of the Paleocene Glires</b> . . . . .	<b>99</b>
	Łucja Fostowicz-Frelik	
<b>6</b>	<b>Continuous Spectrum of Lifestyles of Plant-Associated Fungi Under Fluctuating Environments: What Genetic Components Determine the Lifestyle Transition?</b> . . . . .	<b>117</b>
	Kei Hiruma	
<b>7</b>	<b>Genome Evolution of Asexual Organisms and the Paradox of Sex in Eukaryotes</b> . . . . .	<b>133</b>
	Elvira Hörandl, Jens Bast, Alexander Brandt, Stefan Scheu, Christoph Bleidorn, Mathilde Cordellier, Minou Nowrousian, Dominik Begerow, Anja Sturm, Koen Verhoeven, Jens Boenigk, Thomas Friedl, and Micah Dunthorn	

<b>8</b>	<b>On the Origin of Life and Evolution of Living Systems from a World of Biological Membranes</b> . . . . .	169
	Aditya Mittal, Suneyna Bansal, and Anandkumar Madhavjibhai Changani	
<b>9</b>	<b>Orthology: Promises and Challenges</b> . . . . .	203
	Yannis Nevers, Audrey Defosset, and Odile Lecompte	
<b>10</b>	<b>Prehistoric Stone Projectile Points and Technological Convergence</b> . . . . .	229
	Michael J. O'Brien and George R. McGhee	
<b>11</b>	<b>Diversity and Evolution of RNase P</b> . . . . .	255
	Isabell Schencking, Walter Rossmanith, and Roland K. Hartmann	
<b>12</b>	<b>An Unusual Evolutionary Strategy: The Origins, Genetic Repertoire, and Implications of Doubly Uniparental Inheritance of Mitochondrial DNA in Bivalves</b> . . . . .	301
	Donald T. Stewart, Sophie Breton, Emily E. Chase, Brent M. Robicheau, Stefano Bettinazzi, Eric Pante, Noor Youssef, and Manuel A. Garrido-Ramos	
<b>13</b>	<b>The Evolution of the <i>FLOWERING LOCUS T-Like (FTL)</i> Genes in the Goosefoot Subfamily <i>Chenopodioideae</i></b> . . . . .	325
	Helena Štorchová	
<b>14</b>	<b>DDE Transposon as Public Goods</b> . . . . .	337
	Louis Tsakou-Ngouafo, Célia Vicari, Laura Helou, Vivek Keshri, Sabyasachi Das, Yves Bigot, and Pierre Pontarotti	
<b>15</b>	<b>Evolution of Milk Oligosaccharides of Carnivora and Artiodactyla: Significance of the Ratio of Oligosaccharides to Lactose in Milk</b> . . . . .	359
	Tadasu Urashima, Yuri Mineguchi, Kenji Fukuda, Katherine Whitehouse-Tedd, and Olav T. Oftedal	
<b>16</b>	<b>Making Sense of Noise</b> . . . . .	379
	Shu-Ting You and Jun-Yi Leu	

# Chapter 1

## Genetic and Morphological Differentiation of Common Toads in the Alps and the Apennines



Jan W. Arntzen, Wouter de Vries, Daniele Canestrelli,  
and Iñigo Martínez-Solano

**Abstract** Using a panel of 31 diagnostic nuclear SNP markers in 56 toad populations from the southern Alps and the northern and central Apennines, we document that the range of the spined toad, *Bufo spinosus*, extends along the Mediterranean coast from France into the northwest of Italy. This species, and the common toad *B. bufo*, engage in a unimodal hybrid zone, with *B. spinosus* at the higher and *B. bufo* at the lower altitudes of the Ligurian Alps. The width of 24.0 km observed in the Italian section of the long hybrid zone is narrower than documented in France (ca. 50 km). Using six mitochondrial SNP markers, we resolve several haplotype groups, with approximate distributions as ‘bufo-e2’ across the French–Italian border region, ‘bufo-e3’ in the Italian Alps, ‘bufo-e6’ in the Apennines and ‘spinosus’ with a single occurrence in Italy. We did not observe the northern European ‘bufo-e1’ haplogroup. The bufo-e2 haplogroup is frequently observed in *B. spinosus*, thereby confirming a marked nuclear/mitochondrial discordance. The bufo-e3 and e6 haplogroups correspond to separate nuclear genetic lineages of *B. bufo* in the Alps and the Apennines, respectively. These groups engage in a wide zone of intergradation (109.3 km). *Bufo bufo* populations from the Po valley remain to be studied, but presumably fall with the northern Alpine group. With data from the southeast of France as a reference, we show that Italian *B. bufo*, in particular those from the

---

**Electronic supplementary material** The online version of this chapter ([https://doi.org/10.1007/978-3-030-57246-4\\_1](https://doi.org/10.1007/978-3-030-57246-4_1)) contains supplementary material, which is available to authorized users.

---

J. W. Arntzen (✉) · I. Martínez-Solano  
Naturalis Biodiversity Center, P.O. Box 9517, 2300 RA Leiden, The Netherlands  
e-mail: [pim.arntzen@naturalis.nl](mailto:pim.arntzen@naturalis.nl)

W. de Vries  
Asociación Ambor, Ctra. Constantina—Pedroso 1, 41450 Constantina, Spain

D. Canestrelli  
Department of Ecological and Biological Science, Largo Dell’Università S.N.C, 01100 Viterbo, Italy

I. Martínez-Solano  
Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales, CSIC, c/ José Gutiérrez Abascal, 2, 28006 Madrid, Spain



Apennine lineage, are morphologically more similar to *B. spinosus* than to *B. bufo*. A two-species distribution model developed for France on the basis of climate variables was extrapolated over the entire *B. spinosus*–*B. bufo* range. The result suggests that large parts of the central and eastern Mediterranean have climates that in southern France and the western Mediterranean region support the presence of *B. spinosus*. This clarifies taxonomic confusion regarding species assignment of common toads from southern European peninsulas and hints at possible processes of convergent adaptive morphological evolution in these areas.

## 1.1 Introduction

Many widespread taxa have, over the last decades, turned out to be conglomerates of distinct, but closely related and morphologically similar species, also known as ‘cryptic species’. Amphibian examples include ‘*Triton cristatus*’, ‘*Triton taeniatus*’ and ‘*Rana esculenta*’ in the Old World and ‘*Rana pipiens*’ in the New World. These taxa are currently considered species groups within the genera *Triturus*, *Lissotriton*, *Pelophylax* and *Lithobates* (Frost 2019). Mostly because of their ubiquity and availability, these very species have been popular for teaching purposes and for morphological, physiological and behavioural research (e.g., Rusconi 1821; Ecker 1864; Ecker et al. 1899; Donaldson 1908; Tinbergen and Ter Pelkwijk 1938). To illustrate the point, only six taxa (including all four listed above) were instrumental to the seven Nobel Prize winning studies and 12 laureates that made use of amphibian models (Burggren and Warburton 2007). It remains, however, unclear which of the currently recognized species were actually employed in these studies and how taxonomic affiliation might influence the reproducibility of results in these and other investigations. This highlights the importance of taxonomic knowledge linked to a robust phylogenetic framework. In this respect, molecular systematics has decisively contributed to delineate taxa and resolve phylogenetic relationships in groups comprising morphologically similar species, which has led to rapid progress in the resolution of terminal branches in the tree of life, especially with the advent of high-throughput sequencing technologies (Hinchliff et al. 2015).

One such example of hidden complexity is ‘*Bufo bufo*’ which comprises four species of Eurasian toads. The species are unequivocally diagnosable from genetic markers but difficult to identify from morphology, at least in some regions (Litvinchuk et al. 2008; Recuero et al. 2012; Arntzen et al. 2013, 2016). Within this group the spined toad *Bufo spinosus* Daudin, 1803 from North Africa and south-western Europe and the common toad *Bufo bufo* (Linnaeus, 1758) from western, northern, central and eastern Europe, are genetically deeply differentiated but morphologically cryptic species. The species meet up in a diagonal line across France, from the Atlantic Ocean to the Mediterranean (Arntzen et al. 2018, 2020). They are not sister taxa, yet in the two areas where they were studied in detail, they hybridize in ca. 50 km wide zones (Arntzen et al. 2016; Trujillo et al. 2017; Van Riemsdijk et al. 2019a, b). Nuclear genetic data indicate the presence of *B.*

*spinosus* along the French Mediterranean coast, up to the Italian border, suggesting that the species' range may reach into Italy (Arntzen et al. 2016). We here investigate the distribution and possible genetic admixture of both species over the north of Italy with an eightfold number of nuclear genetic markers (from four to 31). We confirm that *B. spinosus* is present in the northwest of the country (Liguria) and that it engages with *B. bufo* in a narrow hybrid zone. In Italian *B. bufo*, we discern two lineages, distributed along the Alps and in the Apennines, respectively. The southern Apennine lineage of *B. bufo* is morphologically more similar to *B. spinosus* than is the northern Alpine lineage, which hints at processes of convergent morphological evolution driven by climatic factors.

## 1.2 Materials and Methods

Tissue samples were collected under licence as toe tips from adults and the tips of tail fins from larvae for 890 individuals at 56 localities across the north of Italy and adjacent France. Fluorescence-based genotyping (Semagn et al. 2014) was used in the Kompetitive Allele-Specific PCR (KASP) system at the single nucleotide polymorphism (SNP) genotyping facility (LGC genomics, UK) of the Institute of Biology, Leiden University. Primer design, PCR setup and data visualization and recording followed Arntzen et al. (2016) and Van Riemsdijk et al. (2019a). The full panel of SNPs was studied with the exclusion of *banp*, *c10orf2* and *klhl* that lacked species diagnosticity and *egflam* that was strongly out of Hardy–Weinberg equilibrium. Data for individuals with more than five SNPs missing were discarded. For the remainder, missing data amounted to  $N = 470$  (1.7%) over 31 nuclear markers. We further designed and applied a panel of six SNP markers to a subset of the material, with the aim to identify the different mtDNA haplogroups that have been documented for the wider region, namely 'spinosus', 'bufo-e1', 'bufo-e2', 'bufo-e3' and 'bufo-e6' (Garcia-Porta et al. 2012; Arntzen et al. 2017). Individuals with more than one SNP data point missing were excluded. For the remainder, missing data amounted to  $N = 38$  (1.5%). Haplogroup information was compiled for 620 individuals from 33 localities.

We also measured adult toads in the field for two characters that are reported to discriminate between *Bufo bufo* and *B. spinosus* (Arntzen et al. 2013), namely body size (as snout-urostyl length, SU1, measured with callipers at 0.1 mm precision) and parotoid angle (Pa, measured with a protractor on digital imagery, with 0.5 degree precision), all by the same observers (WdeV, JWA). Species identity was determined from nuclear genetic data, and individuals from strongly admixed populations were not analysed. The material studied involved 213 males and 45 females of *B. bufo* from the southeast of France, 207 males and 88 females of *B. spinosus* from the southeast of France and adjacent Italy and 254 males and 41 females of *B. bufo* from the north of Italy. Because of a documented sexual size dimorphism, males and females were analysed separately.

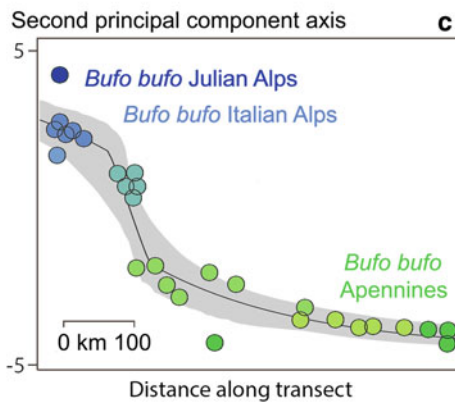
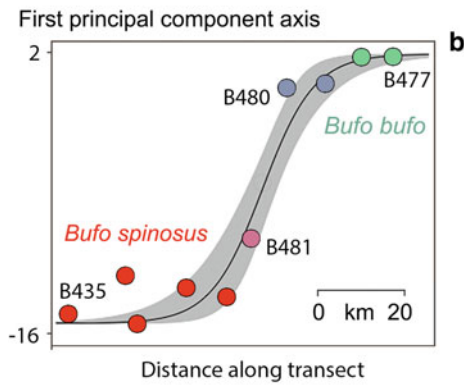
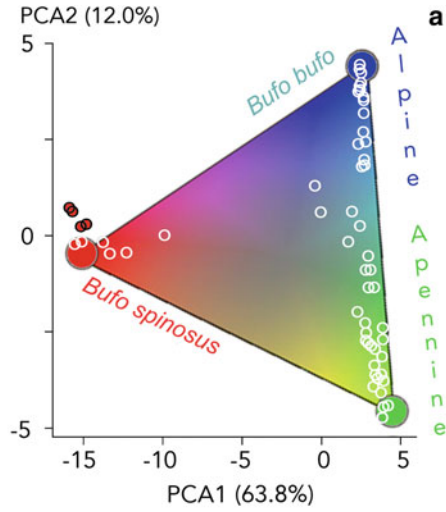
Genetic data were analysed with GenePop (Rousset 2008) and Aegenet (Jombart 2008). Tests for Hardy–Weinberg equilibrium and linkage disequilibrium were evaluated under the Benjamini–Hochberg procedure for multiple comparisons. Heterozygote deficits were observed for locus *ntrk2* in populations B478 and B479, for locus *pdgfr1* in population B391 and for locus *psmg3* in population B479. The signal for linkage disequilibrium was not significant. Molecular genetic clines were investigated with HZAR software as in previous studies (Derryberry et al. 2014; Arntzen et al. 2017; Van Riemsdijk et al. 2019a). Mitochondrial DNA data were analysed in a median-joining network, with PopArt software (Leigh and Bryant 2015). Other statistical analyses were carried out with SPSS 20 (IBM SPSS 2016).

A two-species distribution model was constructed by contrasting published presence data for both species (Arntzen et al. 2020). It describes the contiguous ranges of *B. bufo* and *B. spinosus* from environmental parameters and showed a good fit to the underlying data of 404 genetically investigated toad populations (‘area under the curve statistic’,  $AUC = 0.97 \pm 0.007$ ). The model was re-estimated to only include climate variables (bio01–bio19; Fick and Hijmans 2017) and extrapolated over the combined species ranges including the Mediterranean region. The distribution model was visualized with ILWIS 3.6 (ILWIS 2009).

### 1.3 Results

The first axis of a principal component analysis (PCA) of the nuclear genetic data explains 63.8% of the total variation observed and yields a bimodal distribution, in which *B. spinosus* and *B. bufo* populations are widely separated (Fig. 1.1a). All loci except *psmg3* contribute significantly to the species separation, which reflects the intended diagnosticity of the markers. Three populations are strongly admixed ( $0.2 < F_s < 0.8$ ), with average frequencies of alleles typical for *B. spinosus* ( $F_s$ ) of 0.23 (B480, Ormea),  $F_s = 0.31$  (population B479, Garessio) and  $F_s = 0.73$  (B481, Nava). The reconstructed cline suggests that a major part of the *B. bufo*–*B. spinosus* species transition takes place over a distance of ca. 20 km (Fig. 1.1b). Three moderately admixed *B. spinosus* populations with  $0.8 < F_s < 0.9$  are located at line-in-sight distances of 5.5, 14.0 and 24.5 km to Nava (populations B471, B470 and B381; Fig. 1.2c).

The second PCA axis explains 12.0% of the total observed variation and displays a cluster of *B. spinosus* populations as well as a linear spread of *B. bufo* populations from the Apennine to the Alps (Fig. 1.1a). *Bufo bufo* populations from the Alps and the Apennine regions have mostly positive and negative PCA2 scores, respectively. Significant correlations are observed between  $F_s$  and the scores at the second PCA axis for 14 out of 31 markers (*aimp2*, *b4galt7*, *cwc22*, *dbrr1*, *exon1*, *gatsl2*, *lrrc23*, *med8*, *pigg*, *pdgfr1*, *psmg3*, *rpl3*, *sart3* and *ttc37* with a Spearman correlation coefficient  $r_s$  of  $0.32 < |r_s| < 0.84$ ,  $P < 0.05$ ). A major part of the genetic transition takes place over a distance of ca. 100 km (Fig. 1.1c). Two populations that mark the transition and that are also geographically close to *B. spinosus* are B477 and B478, at



◀**Fig. 1.1** **a** Results of a principal component analysis of nuclear genetic data for *Bufo bufo* and *B. spinosus* from the north of Italy and adjacent France. PCA scores are averaged over populations. The superimposed color triangle aims to link population genetic profiles with geographical position, as in panels band and Fig. 1.2. **b** Sharp nuclear genetic cline described by the first principal component axis. The transect runs over ca. 80 km along the Mediterranean Sea, from pure *B. spinosus* (in red, population B435 in the southwest) to pure *B. bufo* (in green, population B477 in the northeast). Note that the transition is sharpest between populations B481 (locality Nava) and B480 (locality Ormea). The grey shading shows the 95% confidence interval of the cline. For orientation, see Fig. 1.2c. **c** Shallow nuclear genetic cline described by the second principal component axis for Italian *B. bufo*. The transect runs over ca. 600 km, from central Apennine populations (in light green) to west Alpine populations (medium blue, e.g., populations in the western Italian Alps). For orientation, see Fig. 1.2b. A *B. bufo* population from the Julian Alps is shown for comparison (in deep blue)

the northeastern section of the *B. spinosus*–*B. bufo* transect (Fig. 1.2c). These localities approximate the three-way junction for the *B. spinosus*–*B. bufo* Apennine–*B. bufo* Alpine ranges. Numerical details for the clines running in northeastern direction (PCA1, *B. spinosus*–*B. bufo*) and the cline running in southeastern direction (PCA2, Apennine *B. bufo* to Alpine *B. bufo*) are provided in Table 1.1.

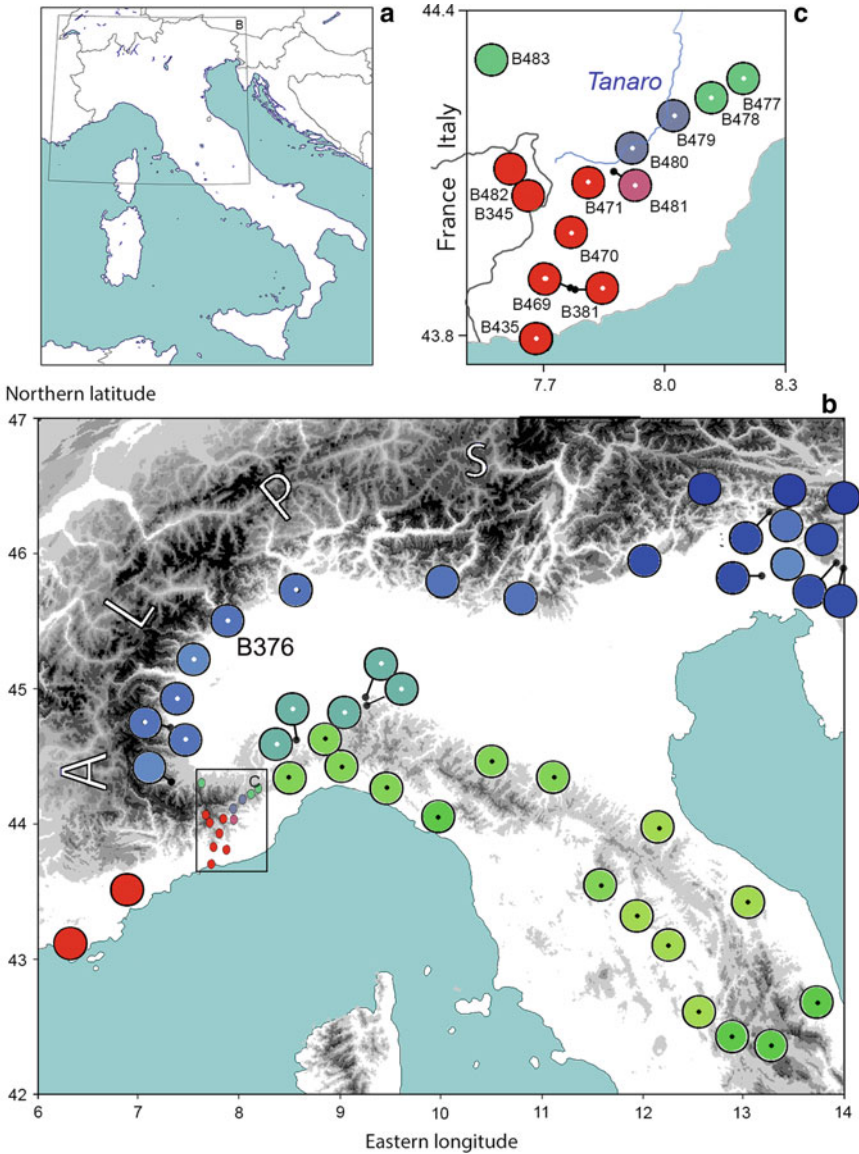
Analysis of mtDNA data revealed a single *spinosus* haplotype in population B381. All other mtDNA haplotypes were typical for *B. bufo*. A total of 187 (30.2%) were of the bufo-e2 type, 145 (23.4%) were of the bufo-e3 type, and 277 (44.7%) were of the bufo-e6 type (Fig. 1.3). The bufo-e1 type was not observed. Finally, for ten individuals (1.6%), haplotypes could not unequivocally be classified because of conflicting diagnostic character states. The spatial distribution of the *B. bufo* haplotype groups is shown in Fig. 1.3.

Bivariate plots of the variables SU1 and Pa confirm the significant morphological differentiation of reference material of both species from France, for males as well as for females. In a classifying discriminant analysis, the far majority of Italian *B. bufo* (81%) is not assigned to its own species but to *B. spinosus*, as illustrated by strongly overlapping ellipses in Fig. 1.4.

The two-species distribution model takes the following logistic equation:  $P_b = 1/(1 + \exp(0.0453 * \text{bio01} - 0.559 * \text{bio02} + 1.423 * \text{bio03} + 0.00844 * \text{bio04} - 0.0284 * \text{bio06} - 0.0237 * \text{bio09} - 51.278))$ , in which  $P_b$  is the probability for the presence of *B. bufo* at the locality investigated, on a zero to unity scale. The model fit in France is  $\text{AUC} = 0.91 \pm 0.015$ . For a spatial representation of the model, see Fig. 1.5.

## 1.4 Discussion

Our increased population and genetic sampling confirms that *B. spinosus* is present in the northwest of Italy and that it engages with *B. bufo* in a narrow hybrid zone in Liguria. The inferred cline width (24 km) is narrower than observed in France (ca. 50 km in the northwest as well as in the southeast of the country (Van Riemsdijk



**Fig. 1.2** Differential nuclear genetic composition of toad populations (spined toad, *B. spinosus* and common toad, *Bufo bufo*) in the north of Italy. **a** Italy, with the research area boxed. **b** Studied populations coloured after their position in the PCA-plot of Fig. 1.1, with *B. spinosus* in red and *B. bufo* in blue and green. Altitudes are from <500 m a.s.l (white) to >2500 m a.s.l. (black shading), with increments of 500 m. **c** Detailed presentation of the *B. spinosus*-*B. bufo* transition in the Ligurian Alps. For sample sizes and locality information, see Supplementary Information. Populations included in the transects of Fig. 1.1b, c are marked by a black or white central dot. The transects arbitrarily start at population B435 to run in northeasterly direction and at population B376 to run in southeasterly direction

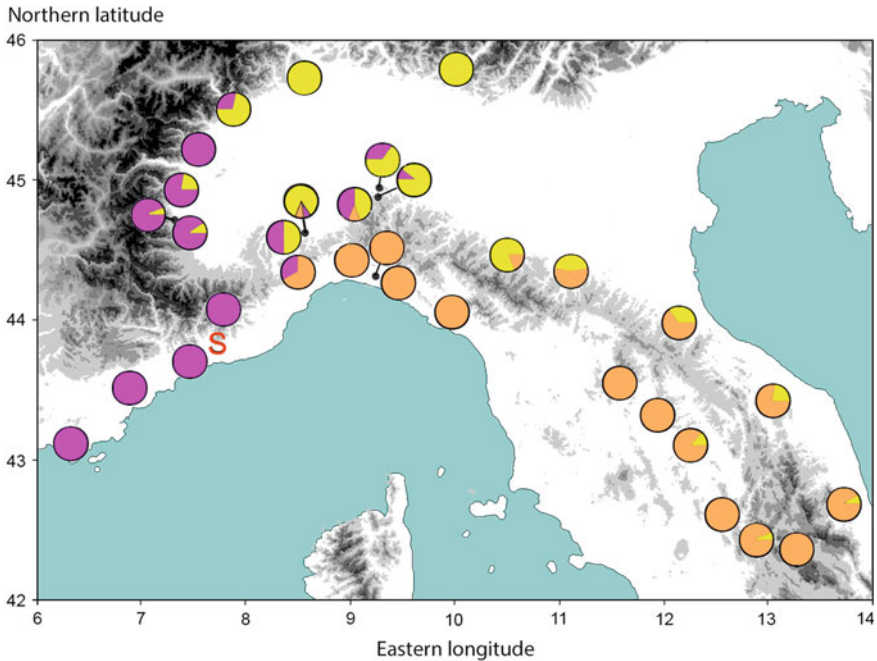
**Table 1.1** Parameter estimates for the maximum-likelihood geographical clines shown in Fig. 1.1

Cline position is relative to selected populations (see below) and width is 1/maximum slope. Confidence intervals are based upon the 2log-likelihood unit support limits. Delta ( $\delta$ ) and tau ( $\tau$ ) are the shape parameters for the tails, with fitting at right (R) or no fitting (N). Pmin and Pmax are the estimated PCA scores at either end of the transect and are fixed to their empirical values

Character studied	Principal component axis		
	First	Second	
Taxa involved	<i>B. spinosus</i> – <i>B. bufo</i>		
Cline shown in Figure	1.1b	1.1c	
Approximate spatial orientation	Northeast	Southeast	
Starting position	B435	B376	
Model type	FixN	FixR	
Position (km)	44.89	123.38	
95% Confidence interval	Min	40.71	108.86
	Max	48.55	143.52
Width (km)	24.01	109.28	
95% Confidence interval	Min	14.77	32.21
	Max	38.13	223.12
Right tail fitting	$\delta$		22.54
	$\tau$		0.181
P, PCA scores at either end	Min	–15.304	–2.594
	Max	1.906	4.365

et al. 2019a, b), suggesting stronger selection against hybrids in Italy. Whether this selection is the result of intrinsic or extrinsic factors, or that both elements contribute to maintaining this tension zone, remains an open question. With respect to climate and topography, Arntzen et al. (2020) highlighted the role of major rivers as barriers to dispersal. However, populations B479 and B480 are actually breeding in the margins of the river Tanaro (Fig. 1.2c) and possess similar genetic profiles, suggesting that the river supports a panmictic population and acts as a route for dispersal along its upper stretches, rather than as a barrier. This genetically admixed riverine population is likely to extend further upstream, in which case the centre of the hybrid zone could be pinpointed to in between the villages Nava and Ponte di Nava, at 900–810 m a.s.l., reducing hybrid zone width to just a few kilometres.

We also found considerable variation in Italian *B. bufo*, with two major nuclear lineages distributed along the Alps and in the Apennines, respectively. Our SNP panel was developed to discriminate *B. spinosus* and *B. bufo* in France, where haplogroups bufo-e1 and bufo-e2 are present. The presence of genetically differentiated lineages bufo-e3, e4 and e6 in Italy may have affected the discriminant performance of our SNP panel, but the results, based on large sample sizes, seem robust to possible ascertainment bias. At any rate, the differentiation observed in Italian *B. bufo* suggests



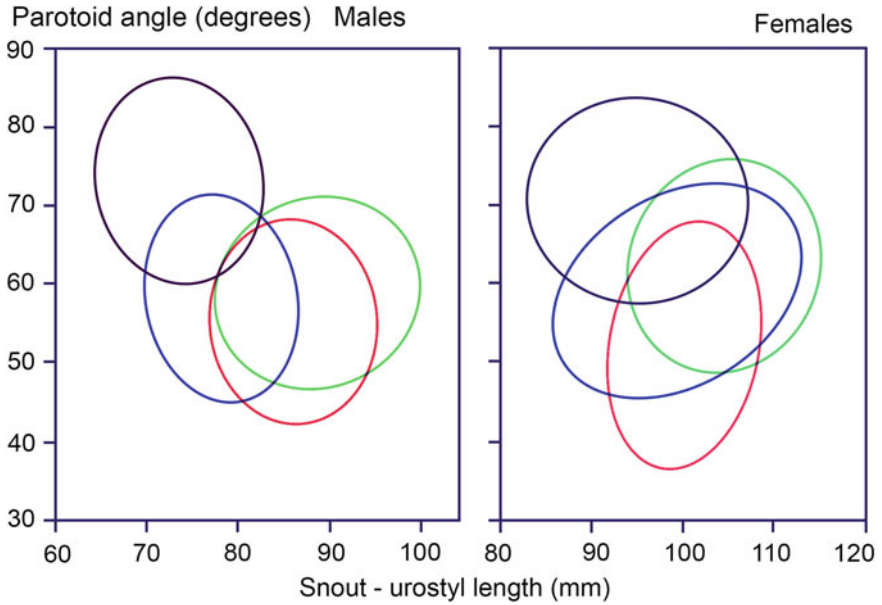
**Fig. 1.3** Distribution of *Bufo bufo* mitochondrial DNA haplotypes over 33 localities in the southern Alps and the Apennines, with bufo-e2 in purple, bufo-e3 in yellow and bufo-e6 in orange. The occurrence of a single rare spinosus haplotype is shown by a red ‘S’. Note that the bufo-e1 type was not found and that all but one of the *B. spinosus* along the Mediterranean coast (Fig. 1.2) carry the bufo-e2 haplotype. For sample sizes and locality information, see Supplementary Information

that the Apennine peninsula has been an important glacial refugium for the species. Future studies should include samples from putative southern refugia (Calabria, Sicily), where phylogeographic studies of other amphibian taxa have revealed deeply diverged lineages (see for instance Canestrelli et al. 2015).

Genetic differentiation between Alpine and Apennine lineages in the Italian peninsula is a common phylogeographic pattern, previously described in a number of taxa, including mammals (Chiocchio et al. 2019), but specially amphibians (Bisconti et al. 2018; Chiocchio et al. 2017; Dufresnes et al. 2018). This hints at common historical events, causing similar responses in biotic communities, that can be inferred from their genomic signatures. Both lineages admix broadly in the upper reaches of the Po plain, with cline width exceeding 100 km being consistent with a lack of barriers to hybridization.

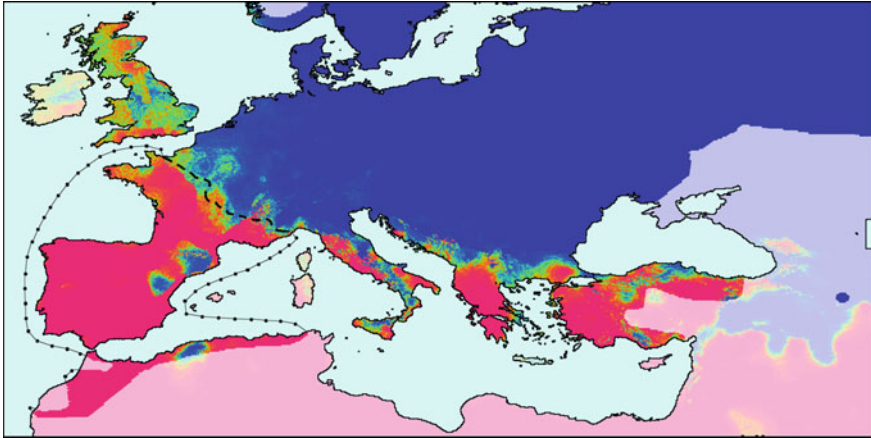
Another important finding of our study is the fact that Italian *B. bufo* are morphologically similar to *B. spinosus* from the south of France and different from conspecifics from France. This reflects the—now abandoned— notion of *B. spinosus* as a circum-Mediterranean subspecies (Sinsch et al. 2009) and suggests convergent evolution of Mediterranean *B. bufo* to the *B. spinosus* phenotype. Alternatively, the





**Fig. 1.4** Bivariate plot of the two main morphological characters (paratoid angle and snout-urostyl length) that discriminate *Bufo bufo* (black) and *B. spinosus* (red) in the southeast of France. Data are summarized by ellipses that represent the mean and one standard deviation, for males (left panel) and for females (right panel). Note that for both sexes *B. bufo* from Italy (in blue and green) are morphologically more similar to *B. spinosus* than to conspecific French *B. bufo*, irrespective of their origin from the Italian Alps (blue) or the Apennines (green)

widespread northern European *B. bufo* lineage (represented by the e1-haplotype) lost some of the *B. spinosus* features. Thus, our results clarify taxonomic confusion regarding species assignment of common toads from southern European peninsulas and hint at possible processes of convergent adaptive morphological evolution in these areas that could be further explored with a combination of experimental and genome-wide association studies.



**Fig. 1.5** Climate-based two-species distribution model for the spined toad, *Bufo spinosus* (predicted range in red) and the common toad, *B. bufo* (predicted range in blue). Green colours represent intermediate predictions. Shaded areas are outside the range of both species (Agasyan et al. 2009) and the documented range of *B. spinosus* over the southwest of France, the Iberian Peninsula and the Maghreb does not extend across the dotted line. The species contact zone across France and the northwest of Italy is shown by an interrupted line. The distribution model is derived from genetically secured records for France (for details see Arntzen 2019; Arntzen et al. 2020) and is here extrapolated over Europe and adjacent parts of Africa and Asia. Note that large areas of the central Mediterranean (i.e., Italy) and the eastern Mediterranean (i.e., the Balkans and Turkey), actually inhabited by *B. bufo*, have a reconstructed climatic profile similar to that of the *B. spinosus* range. For the situation in Great Britain see Arntzen (2019)

**Acknowledgements** We thank the late Annie Zuiderwijk for assistance in the field and Onno Schaap for running the SNP-line.

## References

- Agasyan A, Avisi A, Tuniyev B, Crnobrnja-Isailovic J, Lymberakis P, Andr n C, Kaya U (2009) *Bufo bufo*. The IUCN red list of threatened species 2009. <https://doi.org/10.2305/IUCN.UK.2009.RLTS.T54596A11159939.en>
- Arntzen JW (2019) An amphibian species pushed out of Britain by a moving hybrid zone. *Mol Ecol* 28:5145–5154. <https://doi.org/10.1111/mec.15285>
- Arntzen JW, Canestrelli D, Mart n z-Solano I (2020) Environmental correlates of the European common toad hybrid zone. *Contrib Zool* 89:270–281. Advance article available at <https://doi.org/10.1163/18759866-bja10001>
- Arntzen JW, de Vries W, Canestrelli D, Mart n z-Solano I (2017) Hybrid zone formation and contrasting outcomes of secondary contact over transects in common toads. *Mol Ecol* 26:5663–5675. <https://doi.org/10.1111/mec.14273>
- Arntzen JW, McAtear J, But t R, Mart n z-Solano I (2018) A common toad hybrid zone that runs from the Atlantic to the Mediterranean. *Amphibia-Reptilia* 39:41–50. <https://doi.org/10.1163/15685381-00003145>

- Arntzen JW, McAtear J, Recuero E, Ziermann JM, Ohler A, van Alphen J, Martínez- Solano I (2013) Morphological and genetic differentiation of *Bufo* toads: two cryptic species in Western Europe (Anura, Bufonidae). *Contrib Zool* 82:147–169. <https://doi.org/10.1163/18759866-08204001>
- Arntzen JW, Trujillo T, Butôt R, Vrieling K, Schaap OD, Gutiérrez-Rodriguez J, Martínez-Solano I (2016) Concordant morphological and molecular clines in a contact zone of the common and spined toad (*Bufo bufo* and *B. spinosus*) in the northwest of France. *Front Zool* 13:1–12. <https://doi.org/10.1186/s12983-016-0184-7>
- Bisconti R, Porretta D, Arduino P, Nascetti G, Canestrelli D (2018) Hybridization and extensive mitochondrial introgression among fire salamanders in peninsular Italy. *Sci Rep* 8:13187. <https://doi.org/10.1038/s41598-018-31535-x>
- Burggren WW, Warburton S (2007) Amphibians as animal models for laboratory research in physiology. *ILAR J* 48:260–269. <https://doi.org/10.1093/ilar.48.3.260>
- Canestrelli D, Bisconti R, Sacco F, Nascetti G (2015) What triggers the rising of an intraspecific biodiversity hotspot? Hints from the agile frog. *Sci Rep* 4:5042. <https://doi.org/10.1038/srep05042>
- Chiocchio A, Bisconti R, Zampiglia M, Nascetti G, Canestrelli D (2017) Quaternary history, population genetic structure and diversity of the cold-adapted Alpine newt *Ichthyosaura alpestris* in peninsular Italy. *Sci Rep* 7:2955. <https://doi.org/10.1038/s41598-017-03116-x>
- Chiocchio A, Colangelo P, Aloise G, Amori G, Bertolino S, Bisconti R, Castiglia R (2019) Population genetic structure of the bank vole *Myodes glareolus* within its glacial refugium in peninsular Italy. *J Zool Syst Evol Res* 57:959–969. <https://doi.org/10.1111/jzs.12289>
- Derryberry EP, Derryberry GE, Maley JM, Brumfield RT (2014) HZAR: hybrid zone analysis using an R software package. *Mol Ecol Resour* 14:652–663. <https://doi.org/10.1111/1755-0998.12209>
- Donaldson HH (1908) The nervous system of the American leopard frog, *Rana pipiens*, compared with that of the European frogs, *Rana esculenta* and *Rana temporaria* (Fusca). *J Comp Neurol Psychol* 18:121–149. <https://doi.org/10.1002/cne.920180203>
- Dufresnes C, Mazepa G, Rodrigues N, Brelsford A, Litvinchuk SN, Sermier R, Lavanchy G, Betto-Colliard C, Blaser O, Borzée A, Cavoto E, Fabre G, Ghali K, Gossen C, Horn A, Leuenberger J, Phillips BC, Saunders PA, Savary R, Maddalena T, Stöck M, Dubey S, Canestrelli D, Jeffries DL (2018) Genomic evidence for cryptic speciation in tree frogs from the Apennine peninsula, with description of *Hyla perrini* sp. nov. *Front Ecol Evol* 6. <https://doi.org/10.3389/fevo.2018.00144>
- Ecker A (1864) Die Anatomie des Frosches: Ein Handbuch für Physiologen, Ärzte und Studierende. Friedrich Vieweg und Sohn, Braunschweig, Germany. <https://doi.org/10.5962/bhl.title.5512>
- Ecker A, Wiedersheim R, Gaupp EWT (1899) Anatomie des Frosches. Friedrich Vieweg und Sohn, Braunschweig, Germany
- Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 37:4302–4315. <https://doi.org/10.1002/joc.5086>
- Frost DR (2019) Amphibian species of the World: an Online Reference. Version 6.0. American Museum of Natural History, New York, USA. <https://research.amnh.org/herpetology/amphibia/index.html>
- García-Porta J, Litvinchuk SN, Crochet PA, Romano A, Geniez PH, Lo-Valvo M, Lymberakis P, Carranza S (2012) Molecular phylogenetics and historical biogeography of the west-palaearctic common toads (*Bufo bufo* species complex). *Mol Phylogenet Evol* 63:113–130. <https://doi.org/10.1016/j.ympev.2011.12.019>
- Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci* 112:12764–12769. <https://doi.org/10.1073/pnas.1423041112>
- IBM SPSS (2016) Statistical package for the social sciences. SPSS Inc., Chicago, USA
- ILWIS (2009) Integrated land and water information system (ILWIS). Open software version 3.6. Enschede, The Netherlands, ITC
- Jombart T (2008) ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Leigh JW, Bryant D (2015) PopArt: full-feature software for haplotype network construction. *Methods Ecol Evol* 6:1110–1116. <https://doi.org/10.1111/2041-210X.12410>

- Litvinchuk SN, Borkin LJ, Skorinov DV, Rosanov JM (2008) A new species of common toads from the Talysh mountains, south-eastern Caucasus: genome size, allozyme, and morphological evidences. *Russ J Herpetol* 15:19–43
- Recuero E, Canestrelli D, Vörös J, Szabó K, Poyarkov NA, Arntzen JW, Crnobrnja-Isailovic J, Kidov AA, Cogălniceanu D, Caputo FP, Nascetti G (2012) Multilocus species tree analyses resolve the radiation of the widespread *Bufo bufo* species group (Anura, Bufonidae). *Mol Phylogenet Evol* 62:71–86. <https://doi.org/10.1016/j.ympev.2011.09.008>
- Rousset F (2008) GenePop'007: a complete re-implementation of the GenePop software for Windows and Linux. *Mol Ecol Resour* 8:103–106. <https://doi.org/10.1111/j.1471-8286.2007.01931.x>
- Rusconi M (1821) *Amours des Salamandres aquatiques*. Paolo Emilio Giusti, Milan
- Semagn K, Babu R, Hearne S, Olsen M (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement. *Mol Breeding* 33:1–14. <https://doi.org/10.1007/s11032-013-9917-x>
- Sinsch U, Schneider H, Tarkhishvili D (2009) *Bufo bufo* Superspezies – Erdkröten-Artenkreis – taxon *bufo* (Linnaeus, 1758) – Erdkröte – taxon *gredosicola* L. Müller und Hellmich 1935 – Gredoserdkröte – taxon *spinosus* Daudin, 1803 – Riesenerdkröte – taxon *verrucosissimus* (Pallas, 1811) – Kolchische Erdkröte. pp 191–337. In: K. Grossenbacher: *Handbuch der Reptilien und Amphibien Europas*, Band 5/1 Anuren. Aula-Verlag, Wiebelsheim, Germany
- Tinbergen N, Ter Pelkewijk JJ (1938) De kleine watersalamander. *De Levende Natuur* 43:232–237
- Trujillo T, Gutiérrez-Rodríguez J, Arntzen JW, Martínez-Solano I (2017) Morphological and molecular data to describe a hybrid population of the Common toad (*Bufo bufo*) and the Spined toad (*Bufo spinosus*) in western France. *Contrib Zool* 86:1–9. <https://doi.org/10.1163/18759866-08601001>
- Van Riemsdijk I, Arntzen JW, Butlin RK, Bucchiarelli G, McCartney-Melstad E, Rafajlovic M, Wielstra B (2019) Spatial variation in introgression along the common toad hybrid zone. In: van Riemsdijk I (ed) *Hybrid zone dynamics in amphibians*, pp 81–91. Ph.D. thesis. Leiden University, Leiden, The Netherlands. ISBN: 978-94-6380-475-2
- Van Riemsdijk I, Butlin RK, Wielstra B, Arntzen JW (2019) Testing an hypothesis of hybrid zone movement for toads in France. *Mol Ecol* 28:1070–1083. <https://doi.org/10.1111/mec.15005>

# Chapter 2

## Molecular Phenotypes as Key Intermediates in Mapping Genotypes to Fitness



Aditya Ballal, Constantin D. Malliaris, and Alexandre V. Morozov

**Abstract** We argue that the staggering complexity of the relationship between the organism's genomic sequence and its evolutionary success, as measured by organismal fitness, can become more tractable when viewed through the lens of phenotypic features. These phenotypic features can refer to molecular properties of a protein or protein complex or represent morphological characteristics such as body size and beak shape in birds. Using protein evolution as an example, we demonstrate that it is possible to describe phenotypic landscapes, in which every protein sequence is associated with a phenotypic value such as free energy of protein folding, using compact and interpretable models that can be learned from relatively small-scale datasets. The predicted phenotypic values then serve as explicit inputs to a model of organismal fitness, with the functional form of the fitness function given by biophysical considerations or learned from evolutionary data (fitness measurements for a collection of genotypes). Thus, instead of being a high-dimensional function of genotypes, fitness becomes a low-dimensional function of one or several phenotypes, making it much easier to visualize fitness landscapes and analyze their properties. Moreover, evolutionary dynamics on such landscapes can be decomposed into generation of phenotypic differences by mutations and subsequent motion on the low-dimensional fitness landscape. This two-tiered approach, made possible by recent advances in high-throughput molecular biology, may hold a key to better understanding of the evolutionary processes that have shaped, and continue to shape, all life on Earth.

### 2.1 Introduction

The ability of a cell to survive, grow, and divide depends on the concerted action of thousands of proteins and other biomolecular machines inside it. Biological functions of most proteins emerge directly from their physico-chemical properties, which are determined by protein structure: complicated mutual arrangements of protein amino

---

A. Ballal · C. D. Malliaris · A. V. Morozov (✉)

Department of Physics & Astronomy and Center for Quantitative Biology,  
Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA  
e-mail: [morozov@physics.rutgers.edu](mailto:morozov@physics.rutgers.edu)

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_2](https://doi.org/10.1007/978-3-030-57246-4_2)

acids (aa) in three-dimensional space. Protein structures are not static but rather change in time in complicated ways (Shaw et al. 2010), due to both thermal fluctuations and larger-scale conformational transitions such as those responsible for allosteric control (regulation of an enzyme by binding an effector molecule at a site other than the enzyme's active site) (Monod et al. 1965) or signal transduction (Gardino et al. 2009). On the evolutionary level, changing one amino acid into another as a result of a random single-nucleotide mutation in the protein coding sequence may alter or abolish its function, which may in turn have profound consequences for the evolutionary fitness of the cell. In unicellular organisms, fitness is related to the cell's ability to survive and divide (Crow and Kimura 1970; Gillespie 2004), as a result, fitness is often equated with growth rates in bacterial or yeast populations.

Although single-nucleotide mutations constitute an important mechanism for generating variation in an evolving population, they are not the only ones: insertions and deletions as well as homologous recombination have the potential to alter protein's sequence and therefore affect its structure and dynamics. Thus, various stochastic mechanisms generate genomic mutations in a population of organisms, providing, as one of the consequences, variation in the functional properties of the proteins in each individual. Since fitness is determined in part by protein function (which in turn is a direct consequence of protein structure and dynamics), it is this variation in the protein functional properties that provides raw material for the natural selection to act upon. In other words, fitness can be expressed more naturally as a function of protein and other molecular phenotypes rather than modeled directly as a function of genomic sequences.

Arguably, this progression from protein coding sequences to molecular phenotypes to organismal fitness is rather protein-centric—for example, it appears to neglect the fact that evolution often acts through changes in protein regulation rather than protein sequence (Wray 2007). However, any functional element in the genome can in principle be treated using the same line of reasoning: relevant molecular phenotypes can be identified and studies of the effect of these phenotypes on organismal fitness can be decoupled from the question of how the observed distribution of molecular phenotypes is generated by the underlying mutational mechanisms. For example, in the evolution of *cis* regulation, free energies of transcription factor-DNA binding or transcription factor on and off rates provide natural choices of the molecular phenotypes to focus on. Thus the “genotype→phenotype→fitness” paradigm (as distinct from direct mapping of genotypes onto fitness) constitutes a universal framework for modeling evolutionary dynamics in diverse populations of organisms.

In this review, we will describe some of the recent progress in identifying, experimentally exploring and modeling molecular phenotypes and their contributions to fitness. Using representative examples of both theoretical work and large-scale experimental studies designed to probe protein function and evolutionary consequences of protein mutations, we will demonstrate substantial conceptual and practical advantages of building evolutionary models that explicitly formulate fitness as a function of one or several molecular phenotypes. We will argue that such an approach may provide the key advance needed to connect a wealth of mutational data available from large-scale genomic studies with the well-established mathematical framework

of population genetics, which studies the effects of mutations, selection, and genetic drift in evolving populations (Crow and Kimura 1970; Kimura 1983; Gillespie 2004; Ewens 2004; Hartl and Clark 2007; Wakeley 2005).

## 2.2 Quantitative Description of Phenotypic and Fitness Landscapes

By definition, in protein fitness landscapes, a fitness value is assigned to every protein sequence (Romero and Arnold 2009; Carneiro and Hartl 2010; Canale et al. 2018; Hartman and Tullman-Ercek 2019). Fitness values can be obtained on the basis of evolutionary experiments in which, e.g., bacterial growth rates are measured in a population, with the only source of variation being the sequence of the protein in question. The rest of the genome should be either identical in sequence or follow a known distribution for all protein variants under study. Fitness landscapes are high-dimensional mathematical objects whose properties may be obscured by employing low-dimensional metaphors (Szendro et al. 2013). Protein phenotypic landscapes are the same mathematically as protein fitness landscapes, except that a phenotypic value of interest rather than fitness value is assigned to each sequence. Probing such landscapes requires biophysical and biochemical rather than evolutionary data, which are often easier to collect.

**Mutational mechanisms and the connectivity of protein sequence spaces.** The notion of a protein fitness or phenotypic landscape is intimately related to the idea of protein sequence space (Smith 1970). Protein sequence space is defined as a network in which each protein sequence forms a node; network edges are single amino acid (single-aa) substitutions. Each node on the landscape is assigned a fitness value, which in the absence of evolutionary information is often represented by some biological or physical property of the protein such as its folding stability or the probability to be in a certain conformational state. While this picture of a molecular fitness landscape has played an important role in understanding evolutionary dynamics on the molecular level, it is important to be aware of its principal assumptions and limitations. Indeed, evolution on a fitness landscape is a result of complex interplay between forces of selection, mutation, and genetic drift (stochastic effects that are always present in finite populations due to discreteness of reproductive events (Kimura 1983; Ewens 2004)). Changing mutational parameters or the population size may have a profound effect on how the population evolves, even if the fitness landscape (i.e., selection) remains the same.

Focusing first on mutational mechanisms, we note that the assumption of evolution via single amino acid substitutions which occur at the same mutation rate  $\mu_{\text{aa}}$  is unrealistic without taking the rules of amino acid translation into account. For example, ALA encoded by the GCA codon can mutate into SER encoded by the TCA codon via a single-nucleotide substitution with rate  $\mu_n$ , whereas at least two single-nucleotide mutations are needed to mutate ALA (represented by CGN in the genetic

code) into ILE, which is encoded by ATA, ATC, and ATT. Since single-nucleotide mutation rates are  $\ll 1$  in all organisms, including unicellular organisms such as bacteria and viruses (Drake 1991; Drake et al. 1998; Lynch et al. 2008; Wielgoss et al. 2011),  $\mathcal{O}(\mu_n^2)$  terms can be neglected and as a result the ALA–ILE edge has to be removed from the protein sequence network. Furthermore, in models that work with real genomic data, a single mutation rate  $\mu_n$  should be replaced by more complex models that take into account genome-wide frequencies of the four nucleotides and the chemical distinction between transition and transversion mutational events (Yang 2006). Indeed, it is expected that  $A \leftrightarrow G$  and  $C \leftrightarrow T$  transitions should occur at a higher rate than, e.g.,  $A \leftrightarrow C$  transversions, which necessitate a much more drastic modification of the chemical structure of the DNA base.

Even when all mutational biases are carefully taken into account, mutational moves on the protein sequence network are restricted to single-nucleotide substitutions. Including insertions and deletions (indels), which have been recently argued to play a key role in the expansion of protein structural folds (Light et al. 2013), as well as recombination, whose evolutionary consequences have been debated for decades in the population genetics literature (Crow and Kimura 1965; Eshel and Feldman 1970; Kondrashov 1988), profoundly changes how the nodes in the protein sequence network are connected to one another and how quickly the sequence space can be explored. While indels can be viewed as another type of mutational move, the relative weights of mutational edges are not fixed when recombination is included—rather, they become dependent on the current genetic state of the population. Indeed, in a population without any genetic variation, recombination cannot create new variants at all. In a population with 50% *aa* and 50% *bb* sequences, the frequency of *ab* and *ba* genotypes will be  $0.25r$  on average under a simple model with non-overlapping generations and binomial sampling with replacement, where  $r$  is the recombination rate, defined as the probability of recombination between sequence positions 1 and 2 per generation. However, in a population with 90% *aa* and 10% *bb* sequences, the average frequency of the *ab* and *ba* genotypes will be just  $0.09r$ . It is therefore critically important to be aware of all the mechanisms through which molecular sequences can evolve in a given population or experimental setup.

**Complex mapping from phenotypes to fitness.** Another major caveat has to do with the definition of fitness. In population genetics literature, fitness is defined as the number of progeny per parent over the parent’s lifetime, or the probability of the individual’s survival (Crow and Kimura 1970; Gillespie 2004). It is difficult to relate such complex quantities, which depend on a number of phenotypic characteristics of each organism, to molecular phenotypes that can be easily explored in the laboratory, such as protein binding stability, specificity and affinity of protein-protein and protein-DNA interactions, or enzymatic rates. Therefore, an assumption is often made, either implicitly or explicitly, that fitness is proportional to the molecular characteristic assayed in a given experiment, or described in a given theoretical model. This assumption is often unwarranted since the relationship between a specific molecular phenotype, such as the probability of a protein to be in a folded and functional state, and organismal fitness is typically unknown, and may be nonlinear



and even non-monotonic. For example, in the case of enzymatic reactions, it is conceivable that there is an optimal concentration of the enzyme's product, so that both under- and overproduction of the product molecule are harmful to the cell. Underproduction may deprive the cell of the urgently needed molecular substance. On the other hand, overproduction may lead to various fitness costs including squandering cellular ribosomal capacity to produce superfluous enzymes, depletion of substrate molecules that may have been costly to make and that might be needed in other pathways, and the necessity to dispose of unwanted products. In such cases, it is more appropriate to speak of phenotypic rather than fitness landscapes, unless the experimental setup includes large-scale measurements of growth rates for a library of mutants with known molecular phenotypes, which can elucidate the link between phenotypic characteristics and fitness.

Furthermore, even if the link between a given phenotype and fitness is carefully mapped out experimentally, it is still subject to the laboratory conditions under which the evolution of the population was examined. Indeed, fitness strongly depends on the environmental conditions, including a variety of factors such as temperature, seasonal atmospheric changes, weather conditions, demographic range, and presence of other species which compete for resources with the population of interest. Another closely related phenomenon is frequency-dependent selection, in which fitness of a given phenotype or genotype depends on its frequency in the population (Ayala and Campbell 1974; Levin 1988). Although it is widely accepted that growth rates of a bacterial population strongly depend on and are shaped by the environment, the role of the latter is poorly understood and as a result is often ignored in quantitative models.

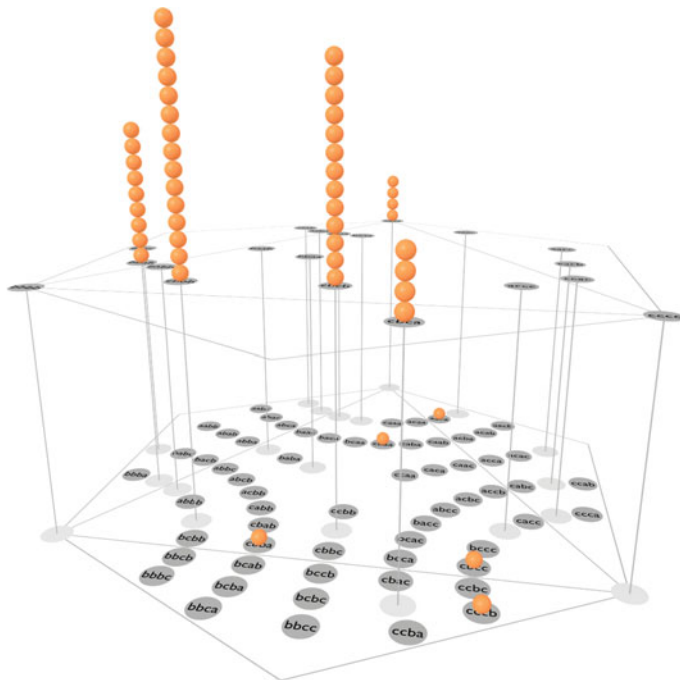
Because of the non-trivial mapping from phenotypes to fitness, it is not always clear how important the observed variation of a given phenotype might be for evolutionary dynamics. In one limiting scenario, observed variations in, e.g., ligand-binding free energy  $\Delta G$  would not affect organismal fitness at all, mapping a rugged "ligand-binding" landscape into a neutral fitness landscape where each protein sequence is assigned the same fitness. Somewhat more realistically, one could construct a "two-plane" fitness landscape in which all low-fitness states are characterized by  $\Delta G$  below a certain threshold, while all high-fitness states correspond to  $\Delta G$  above the threshold. In general, population genetics shows that evolutionary dynamics will be effectively neutral on fitness landscapes with  $N\Delta\mathcal{F} \leq 1$  and  $\Delta\mathcal{F} \leq \mu$ , where  $\Delta\mathcal{F}$  is the difference in fitness between mutationally adjacent sequences,  $\mu$  is the mutation rate, and  $N$  is the effective population size (Crow and Kimura 1970; Gillespie 2004). Thus, it is especially important to know which phenotypic changes can give rise to selective forces that are strong enough not to be masked by genetic drift (stochastic effects in finite-size populations) and mutational effects.

**Epistasis and landscape structure in molecular evolution.** The above arguments indicate that the relevant features of a fitness landscape and, in particular, the distribution of fitness differences between pairs of nodes connected by a single mutational move do not have absolute meaning but rather must be measured against the characteristic scales of the other two evolutionary forces: mutation and genetic drift. In

other words, a landscape would look smooth to a small population buffeted by strong stochastic forces of genetic drift, which will force the population to leave local fitness peaks and enable it to cross fitness valleys. The same landscape would appear much more structured to a large population, whose weak stochastic forces would enable it to localize onto the nearest basin of attraction more efficiently, while making it more difficult to escape such local features.

Nonetheless, even though the landscape structure is not absolute, it plays a key role in classifying fitness landscapes and understanding the types of evolutionary dynamics they can support. In many protein phenotypic landscapes, a natural separation of scales arises: most protein sequences correspond to the “null” phenotype because they are unable to adopt a well-defined fold that would enable them to carry out their biological function, such as binding a ligand or undergoing a conformational change as part of a signal transduction pathway. Only a small fraction of all possible sequences would yield measurable phenotypes, which may have non-trivial patterns of interconnectivity (e.g., isolated islands vs. a single cluster which contains all viable sequences) and exhibit additional features such as local maxima or basins of attraction. The simplest landscape of this kind is a “two-plane” landscape in which each sequence is assigned either a low-fitness (or null phenotype) value or a high-fitness (or viable phenotype) value. When the gap between the low- and high-fitness values is substantial, the evolutionary dynamics on such a landscape is largely driven by the connectivity of the sequences on the upper plane and by the number of deleterious vs. neutral mutations available to a given sequence.

There are many ways in which protein landscapes can be characterized quantitatively: the number of local maxima, the distribution of phenotype or fitness differences between mutational neighbors, the average fractions of neutral, deleterious, and beneficial mutations available to a protein sequence, etc. However, very often the features of molecular landscapes are viewed through the lens of epistasis (Starr and Thornton 2016; Miton and Tokuriki 2016; Canale et al. 2018; Hartman and Tullman-Ercek 2019). In the context of protein evolution, the notion of epistasis refers to the fact that as a rule, phenotypic or fitness effects of introducing an amino acid mutation at a given site in the protein sequence will depend on the states of amino acids at the other sites. In the absence of epistasis on a fitness landscape, evolutionary dynamics is highly predictable and the landscape possesses a single global maximum—a sequence with every amino acid in its highest fitness state (Szendro et al. 2013). On the other hand, since on epistatic landscapes effects of mutations may differ in magnitude and even in sign depending on the rest of the sequence, such landscapes exhibit “ruggedness” or “roughness” and may be characterized by multiple fitness peaks (Carneiro and Hartl 2010; Poelwijk et al. 2011; Starr and Thornton 2016). These landscape features make some genotypes less accessible than others and generally restrict the ensemble of evolutionary trajectories to those paths that avoid low-likelihood steps such as crossing fitness valleys (Weinreich et al. 2005, 2006; Poelwijk et al. 2007; Bridgham et al. 2009; Lobkovsky et al. 2011).



**Fig. 2.1** Two-plane fitness landscape. A model of a two-plane fitness landscape for 81 sequences with alphabet size  $A = 3$  and sequence length  $L = 4$ . All protein sequences are connected via single-aa mutations; edges connecting mutational neighbors are omitted for clarity. 20 of the 81 nodes, highlighted with vertical lines, are assigned to the upper fitness plane with  $\mathcal{F} = 10^{100}$  ( $\mathcal{F} = 1$  for sequences in the lower plane, so that plane separation is effectively infinite). Orange spheres represent individuals in a population of size  $N = 50$ . Shown is a randomly chosen steady-state configuration of the population evolved on this landscape using the Wright–Fisher model of population genetics (Ewens 2004), with  $\mu = 0.1$

**Two-plane fitness landscapes: an idealized paradigm.** We illustrate the main ideas outlined above on a simple example of a two-plane fitness landscape (Fig. 2.1). In this idealized model, any sequence is assigned either high or low fitness value, so that the landscape has a pronounced plateau structure. When the separation between the two fitness values is sufficiently large, low-fitness genotypes are not viable and the population is mostly concentrated on the upper plane. Each sequence on the upper plane may undergo neutral or deleterious mutations. With large fitness plane separation, individuals that mutate to the lower plane are effectively removed from the population, creating a one-way flux from fit to unfit genotypes. Eventually the population configuration will reach a steady state in which, in the infinite-population limit, occupancies of low-robustness nodes with more links to the lower plane will be lower than occupancies of the nodes characterized by high robustness (van Nimwegen et al. 1999) (node robustness is defined in terms of the number of deleterious versus neutral moves available to it).

However, protein sequence spaces are so large that the opposite limit may be more biologically relevant: that in which the size of the population is much smaller than the number of high-fitness sequences. In this limit, the population will continue to explore the upper plane, but its population statistics such as the average number of occupied network nodes will no longer change with time. For any given generation, the structure of the population on a two-plane fitness landscape will be determined by the interplay between mutational forces, which are ultimately responsible for the robustness effect, and stochasticity due to the finite population size. One such configuration is shown in Fig. 2.1, although admittedly this example is not as extreme in terms of the ratio of the population size to the number of high-fitness sequences in the protein space as real-world systems might be.

Although the landscape in Fig. 2.1 does not have any structure beyond separation into two fitness states, it does exhibit non-trivial evolutionary properties. For example, it is characterized by widespread epistasis: a single-aa mutation may be neutral or deleterious depending on the rest of the sequence (although for obvious reasons a deleterious mutation can never change its sign, which would have created so-called sign epistasis—a necessary condition for the appearance of local maxima (Poelwijk et al. 2011)). Moreover, the connectivity of the nodes on the upper plane restricts the ensemble of evolutionary trajectories available to the population and together with population-genetic parameters determines the scale of evolutionary times. Therefore, such landscapes can be used to construct simplified models of evolution that are more treatable with currently available computational and analytical tools than more complex systems. At the very least, such models can highlight the connection between robustness and evolvability (Wagner 2008) and serve as a starting point for more detailed descriptions.

## 2.3 Disentangling the Connections Between Genotypes, Phenotypes, and Fitness

**Biophysical phenotypic landscapes for protein folding and binding.** So, is there any hope of understanding exceedingly complex relationships between genotypes, phenotypes, and the observed patterns of evolutionary dynamics in an evolving population? Although the task seems to be daunting, we have several potential advantages on our side. First of all, at least for some phenotypes (quantitative traits), building genotype-to-phenotype maps yields compact, interpretable models, enabling a significant reduction in the dimensionality of the problem. In other words, unlike fitness landscapes, phenotypic landscapes appear to lend themselves more easily to compact descriptions with relatively few parameters that can be learned reliably even from sparse datasets. The protein folding and ligand-binding free energies seem to be in this category since they can be usefully modeled as the sum over all residues which make independent additive contributions to the total (Wells 1990; Serrano et al. 1993; Zhang et al. 1995):

$$E(\sigma) = E^0 + \sum_{i=1}^L \epsilon(\sigma_i), \quad (2.1)$$

where  $E(\sigma)$  is the folding or binding free energy of sequence  $\sigma$ ,  $E^0$  is the folding or binding free energy of an arbitrary reference sequence,  $L$  is the total number of amino acids in the protein, and  $\epsilon(\sigma_i)$  are free energy differences, with respect to the reference sequence, due to amino acid  $\sigma_i$  at position  $i$  in sequence  $\sigma$ . Note that such a landscape, which has  $20^L$  distinct protein sequences (an astronomical number for any realistic protein length), can be fully characterized by measuring or predicting just  $(A - 1)L \epsilon(\sigma_i)$  values and 1 reference value, where  $A$  is the alphabet size.

The “one-body” approximation in Eq. (2.1) is an oversimplification—pairs of residues that are close to each other in the protein structure or that contact each other across a binding interface often form specific interactions such as salt bridges or hydrogen bonds, which should be described by explicit “two-body” terms  $\epsilon(\sigma_i, \sigma_j)$ . Furthermore, as studies of protein double-mutant cycles have shown, even longer-range non-additive couplings are possible, possibly due to protein allostery and rigid-body dynamics of protein domains (Istomin et al. 2008). However, even such models, which require  $\mathcal{O}(A^2L^2)$  fitting parameters (in practice, the number of additional parameters may not be very large because most residue pairs are adequately described by the additive model), represent a vast reduction in dimensionality compared to the exponential complexity of the original protein sequence space.

The above arguments indicate that it may be beneficial to consider phenotypic landscapes in terms of an expansion in which higher-order terms (describing residue pairs, triplets, etc.) are progressively added to the basic model of Eq. (2.1):

$$E(\sigma) = E^0 + \sum_{i=1}^L \epsilon(\sigma_i) + \sum_{i < j}^L \epsilon(\sigma_i, \sigma_j) + \sum_{i < j < k}^L \epsilon(\sigma_i, \sigma_j, \sigma_k) + \dots, \quad (2.2)$$

where  $E$  denotes a phenotype such as a folding or binding free energy, and each  $\epsilon$  is a one-aa, two-aa, three-aa contribution, etc., which depends only on the residues in its argument (the highest order contribution is of order  $L$  and would involve the entire sequence). Note that indices  $i, j, k, \dots$  adopt strictly increasing values in all higher-order terms. Models of this type resemble spin glass or Potts models which have been extensively studied in statistical mechanics (Mezard and Montanari 2009) and machine learning (MacKay 2003), so that numerous computational and theoretical approaches are available for their analysis.

While a detailed discussion of how to fit the model defined by Eq. (2.2) to either simulated or experimentally mapped phenotypic landscapes is beyond the scope of this review, we would like to mention two considerations that occupy a center stage in the model-training process: (i) use of algorithms capable of controlling model sparsity and as a result producing the simplest model compatible with the data; (ii) introduction of constraints into the algorithms that ensure the uniqueness of the

decomposition into terms of increasing order in Eq. (2.1). If the model is left unconstrained, the Eq. (2.1) expansion will be invariant with respect to transformations of the type  $\epsilon(\sigma_i) \rightarrow \epsilon(\sigma_i) - a_i$ ,  $\epsilon(\sigma_j) \rightarrow \epsilon(\sigma_j) - b_j$ ,  $\epsilon(\sigma_i, \sigma_j) \rightarrow \epsilon(\sigma_i, \sigma_j) + a_i + b_j$ , among others. Although these spurious degrees of freedom may affect both rates of convergence and the interpretability of the trained model, they are easily controlled by either setting all  $\epsilon$  terms associated with an arbitrarily chosen  $q$  to zero: e.g.,  $\epsilon(\sigma_i = q) = 0$ ,  $\epsilon(\sigma_i, \sigma_j = q) = \epsilon(\sigma_i = q, \sigma_j) = 0$ ,  $\forall i, j$  for a second-order expansion (Morcos et al. 2011), or simply by imposing a LASSO constraint (Tibshirani 1997) on the fit. LASSO constraints automatically minimize the magnitude of each fitting parameter (Bishop 2006), simultaneously enforcing model sparsity and preventing spurious parameter shifts mentioned above.

**NK fitness model.** A model conceptually similar to Eq. (2.2) was developed by Kauffman (Kauffman and Weinberger 1989; Kauffman 1993) as a description of fitness landscapes with a tunable degree of ruggedness and unpredictability of evolutionary paths. In Kauffman’s NK model, genotypes are typically represented as binary strings of length  $N$ , as opposed to the realistic alphabet size  $A = 4$  for nucleotides and  $A = 20$  for amino acids. Each of the  $N$  sites in the gene (or  $N$  genes in the genome) interacts with  $K$  other sites chosen at random. The fitness of genotype  $\sigma$  is given by

$$\mathcal{F}(\sigma) = \sum_{i=1}^N f_i(\sigma_i, \sigma_{n_1(i)} \dots \sigma_{n_K(i)}), \quad (2.3)$$

where  $n_1(i) \dots n_K(i)$  are interaction partners of site  $i$  and  $\sigma_j = \{\mathbf{A}, \mathbf{B}\}$  is the binary state of site  $j$ . The single-site fitness values  $f_i$  are obtained by sampling from a pre-defined continuous distribution; each combination of  $2^{K+1}$  possible states of the argument corresponds to an independent sample, and the value of  $f_i$  is resampled each time its argument is updated. With  $K = 0$ , the NK landscape becomes fully additive and the model becomes equivalent to that of Eq. (2.1). Because in this limit the landscape has regular structure and a single peak, it is sometimes called the “Mount Fuji” model (Aita et al. 2000). The amount of landscape ruggedness or epistasis can be tuned by increasing  $K$  to its maximum value of  $N - 1$ . When  $K = N - 1$ , all fitness values are uncorrelated and a single mutation may change the fitness completely. Such a model is called the “House of Cards” (Kingman 1978) and is likely unrealistic since fitness values of closely related genotypes are expected to be similar. As can be seen by comparing Eqs. (2.2) and (2.3), each individual term in the Eq. (2.2) expansion is very similar to the NK model of order  $K-1$ , with two main differences: (i) for each two-state site in the NK model, the value of fitness is sampled from a distribution rather than inferred from experimental data and (ii) for each site  $i$ ,  $K$  interaction partners are chosen randomly and only once. These are not substantial differences, however Eq. (2.2) can also be used to generate an artificial phenotypic landscape using very similar sampling strategies.

**Modeling fitness as a function of folding and binding molecular phenotypes.** If a sparse model can be used to describe a phenotypic landscape, why not use the same approach to model fitness directly? We would expect such models to require many more higher-order terms because the relationship between phenotypes and fitness is typically nonlinear. As an example, consider a simple protein fitness model which is based on only one molecular phenotype, the free energy difference between protein folded and unfolded states (for simplicity, we consider proteins with two-state folding kinetics (Creighton 1992; Finkelstein and Ptitsyn 2002)). Numerous recent studies have focused on how proteins evolve under the constraint of maintaining thermodynamic stability (Bloom et al. 2005; DePristo et al. 2005; Bloom et al. 2006; Zeldovich et al. 2007; Bloom et al. 2007a, b, c; Bershtein et al. 2008). As a rule, these models assume that the fitness  $\mathcal{F}$  of the organism is proportional to the probability for the protein to be in the folded state:

$$\mathcal{F}(E_f) = \frac{f_0}{1 + e^{\beta E_f}}, \quad (2.4)$$

where  $f_0$  is the proportionality constant,  $E_f$  is the free energy of folding (i.e.,  $\Delta G$ , the free energy difference between its folded and unfolded states), and  $\beta = 1/k_B T \simeq 1.7$  (kcal/mol) $^{-1}$  is the inverse room temperature ( $k_B$  is the Boltzmann constant). This approach assumes that in thermodynamic equilibrium inside a cell, the fraction of functional proteins in a folded state will be higher if these proteins are more thermodynamically stable. Under the additional assumption that the protein has to adopt its folded state to be functional (which excludes intrinsically disordered proteins (Brown et al. 2011)), the model confers fitness advantages to thermodynamically stable proteins, albeit with diminishing returns as the sigmoid function in Eq. (2.4) saturates at its maximum value. Note that the fitness landscape defined by Eq. (2.4) is characterized by epistasis (although not sign epistasis since the fitness function is monotonic) and by pairwise and higher-order couplings between residues that may be distant in three-dimensional space, simply because of the nonlinear nature of the fitness function.

Some studies simplify Eq. (2.4) further by assuming that the folding free energy  $E_f$  only needs to be below a free energy threshold  $E_f^{\text{threshold}}$ ; all proteins with  $E_f < E_f^{\text{threshold}}$  are functional and have equal fitness. Mathematically,

$$\mathcal{F}(E_f) = f_0 \Theta(E_f^{\text{threshold}} - E_f), \quad (2.5)$$

where  $\Theta$  is the Heaviside step function. This is equivalent to the zero-temperature limit of Eq. 2.4. The resulting fitness landscape is of the two-plane type described above (Fig. 2.1), with the assignment of protein sequences to the upper or lower plane based solely on their folding free energy. Similar models of evolution formulated in terms of protein–DNA free energies of binding have been used to study evolution of transcription factor binding sites (Sengupta et al. 2002; Gerland and Hwa 2002; Berg and Lässig 2003; Berg et al. 2004; Lässig 2007; Mustonen et al. 2008; Haldane et al. 2014).

An obvious extension of the above considerations is a model which depends on two molecular phenotypes: protein folding stability and protein–ligand binding affinity (Manhart and Morozov 2013, 2014, 2015a,b). We focus again on proteins with two-state folding kinetics (Creighton 1992; Finkelstein and Ptitsyn 2002) and assume that such proteins can bind their cognate ligand only when they are folded. Under the thermodynamic equilibrium assumption, valid when protein folding and binding are much faster than characteristic timescales of cell growth and division, the probabilities of the three structural states—folded and bound ( $p_{f,b}$ ), folded and unbound ( $p_{f,ub}$ ), and unfolded and unbound ( $p_{uf,ub}$ )—are given by their respective Boltzmann weights:

State	Free energy	Probability
Folded, bound	$E_f + E_b$	$p_{f,b} = Z^{-1} e^{-\beta(E_f + E_b)}$
Folded, unbound	$E_f$	$p_{f,ub} = Z^{-1} e^{-\beta E_f}$
Unfolded, unbound	0	$p_{uf,ub} = Z^{-1}$

(2.6)

Here,  $\beta$  is the inverse temperature,  $E_f$  is the free energy of folding, and  $E_b = E'_b - \mu$ , where  $E'_b$  is the binding free energy and  $\mu$  is the chemical potential of the ligand. For simplicity, we will refer to  $E_b$  as the binding energy. Note that  $E_f < 0$  for intrinsically stable proteins and  $E_b < 0$  for favorable binding interactions. Finally, the partition function is  $Z = e^{-\beta(E_f + E_b)} + e^{-\beta E_f} + 1$ . The folding and binding energies depend on the amino acid sequence  $\sigma$  through Eq. (2.1), with  $\epsilon$  values either known experimentally for a protein of interest or sampled from “typical” distributions obtained by pooling mutational data for several proteins (Tokuriki et al. 2007).

The fitness landscape is based on two molecular traits  $E_f$  and  $E_b$  which determine the probabilities of the three protein states considered in this model:

$$\mathcal{F}(E_f, E_b) = f_0 (p_{f,b} + f_{ub} p_{f,ub} + f_{ub} f_{uf} p_{uf,ub}), \quad (2.7)$$

where  $f_{ub}, f_{uf} \in [0, 1]$  are the multiplicative fitness penalties for being unbound and unfolded, respectively: the fitness is  $f_{ub}$  if the protein is unbound but folded and  $f_{ub} f_{uf}$  if the protein is both unbound and unfolded. Similar to the single-trait model defined by Eq. (2.4), organismal fitness is assumed to depend linearly on the relative fractions of each of the three structural states in an equilibrium ensemble of protein molecules inside a cell. Surprisingly, despite the fact that the fitness function in Eq. (2.7) is smooth and monotonic, it can give rise to multiple local fitness maxima in the regions where the binding and folding quantitative traits are coupled (Manhart and Morozov 2015a).



We would like to reiterate that in all biophysical models described above, a direct fit of Eq. (2.2) to the fitness landscape is inadvisable because it is likely to result in many higher-order terms, despite the fact that the underlying phenotypic landscapes are represented by the lowest order expansions [Eq. (2.1)]. The situation is somewhat similar to that found in artificial feed-forward neural networks (ANNs), where each neuron receives a sum of weighted inputs from neurons in the previous layer; these sums, called activations, are then used as inputs to nonlinear activation functions such as sigmoids or hyperbolic tangents (Bishop 2006). Similar to ANNs, it may be worthwhile to use the expansion in Eq. (2.2) for phenotypes and either learn parameterized biophysical fitness functions explicitly from evolutionary data or impose their functional form from first principles, as was done, e.g., in Eq. (2.7).

**Alternative approaches to modeling molecular phenotypic landscapes.** Although the functional form of Eq. (2.2) is very flexible and provides a natural framework for fitting phenotypic data, other approaches are possible in which an explicitly defined function such as a Gaussian or a mixture of Gaussians is used to capture key landscape structures, for example, basins of attraction leading to multiple local maxima. In some cases, a fixed amount of noise, typically sampled from a zero-mean normal distribution, may be added to these global functions, usually to mimic landscape roughness in theoretical work (Szendro et al. 2013). The simplest model of this kind, in which each fitness or phenotypic value is an independent random variable sampled from a normal distribution (i.e., there is no global function), is very similar to the NK fitness model in the  $K = N - 1$  limit, and to the random-energy model originally introduced to describe disordered systems in the context of spin-glass physics (Derrida 1980, 1981; Mezard and Montanari 2009).

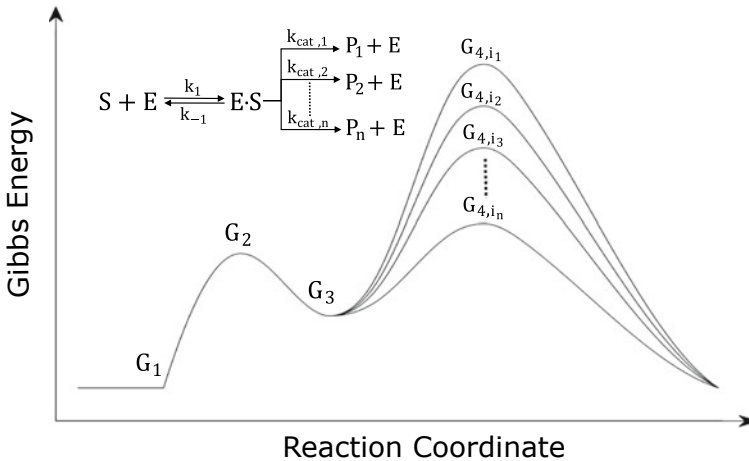
We note that these statistically defined landscapes are typically described by just a few parameters, such as the mean and the standard deviation of the normal distribution in random-energy-type models. These parameters can be easily learned from the data, even with sparse datasets. Although such models may not provide a deterministic fit for each specific value on the landscape, they do capture the landscape’s features in the statistical sense. The situation is similar to describing a small volume of gas molecules by its temperature and pressure, instead of attempting to furnish an exact description of the position and velocity of every gas molecule as a function of time. This idea forms the foundation of statistical mechanics (Landau and Lifshitz 1980) and makes it possible to describe complex macroscopic systems such as gases, liquids, or ferromagnets with a high degree of accuracy. The simplest random-energy-type model can be extended by introducing correlated noise, e.g., by making the parameters of the noise distribution dependent on the mutational neighbors of the current sequence. Correlation structure of fitness landscapes has long attracted attention in the evolutionary biology community (Weinberger 1990; Stadler 1996).

A representative example of the above approach is the “rough Mount Fuji” model (Aita et al. 2000), in which a single genotype  $\sigma_0$  is assigned to be the global maximum and the fitness of genotype  $\sigma$  is given by

$$\mathcal{F}(\sigma) = \eta(\sigma) - \theta d(\sigma, \sigma_0), \quad (2.8)$$

where  $d(\sigma, \sigma_0)$  is the Hamming distance (the number of amino acid or nucleotide substitutions) between sequences  $\sigma$  and  $\sigma_0$ ,  $\theta$  is the parameter which controls the slope of the smooth part of the landscape, and  $\eta(\sigma)$  is a zero-mean random variable sampled independently for each sequence  $\sigma$ . The ruggedness of the landscape is controlled by the ratio of  $\theta$  and the standard deviation of the distribution from which  $\eta(\sigma)$  random variables are sampled. Despite its apparent simplicity, including the assumption of the linear global trend, the rough Mount Fuji model was found to capture essential features of several experimental landscapes (Szendro et al. 2013).

**Biophysical phenotypic landscapes for enzyme kinetics.** Evolution of enzyme kinetics provides another example in which it is advantageous to construct fitness landscapes explicitly in terms of biophysical quantitative traits. Enzymatic activity can be described using Michaelis-Menten free energies which control the rates of various enzyme-mediated reactions: association of the enzyme with the substrate, transformation of the substrate into product, and subsequent release of the product molecule (Bozlee 2007; Nelson 2007) (Fig. 2.2a). Although many natural enzymes are fine-tuned by adaptation to produce a single type of product molecule, this is typically not the case with enzyme libraries created in the laboratory. These enzymes tend to catalyze multiple reactions when they exhibit any activity at all (see Salmon et al. (2015) for a representative example). For an enzyme with sequence  $\sigma$ , the rate of producing product  $i$  is given by



**Fig. 2.2** Michaelis-Menten model of enzyme kinetics. Shown are free energy profiles for converting substrate  $S$  into products  $P_1 \dots P_n$ , catalyzed by the enzyme  $E$ .  $G_1, G_2, G_3, G_{4,i_k}$  are Gibbs free energies at the various stages of the enzymatic reactions, and  $k_{-1}, k_1, k_{cat,i_k}$  are the corresponding reaction rates as shown in the inset (product indices  $i_1 \dots i_n$  are sorted in the decreasing order of  $G_{4,i_k}$ ). Note that  $k_1 = A_1 e^{-\beta(G_2 - G_1)}$  and  $k_{-1} = A_{-1} e^{-\beta(G_2 - G_3)}$ , where  $A_1$  and  $A_{-1}$  are proportionality constants. Furthermore, each reaction rate  $k_{cat,i_k}$  depends on the difference between free energies  $G_{4,i_k}$  and  $G_3$  through Eq. (2.9)

$$k_{\text{cat},i}(\sigma) = B_i e^{-\beta(G_{4,i}(\sigma) - G_3(\sigma))}, \quad (2.9)$$

where  $B_i$  is the reaction rate for product  $i$  in the absence of the free energy barrier and Michaelis-Menten free energies  $G_3$  and  $G_{4,i}$  are defined in Fig. 2.2a.

The case study we shall focus on is the evolution of ring-forming reactions in terpene synthases (TPSs), a major enzyme family whose representatives are found in a variety of plants and insects (Tholl 2006). Cyclic terpenes comprise hundreds of stereochemically complex mono- and polycyclic hydrocarbons; they are involved in pollination, plant and insect predator defense mechanisms, and symbiotic relations. They are also widely used as flavors, fragrances, and medicines; a well-known example of the latter is a naturally occurring anti-malarial drug artemisinin extracted from *Artemisia annua*. Terpenes and terpenoids are the primary constituents of many essential oils in medicinal plants and flowers; examples include  $\alpha$ -bisabolol, a monocyclic sesquiterpene alcohol which forms the basis of a colorless viscous oil from German chamomile, and zingiberene, a monocyclic sesquiterpene that is the predominant constituent of ginger oil.

In order to study the evolutionary emergence and the molecular mechanism of terpene cyclization, Paul O'Maille and collaborators have recently created two enzyme mutant libraries (Salmon et al. 2015; Ballal et al. 2020). One library was constructed to sample natural sequence variation in the background of *A. annua* (E)- $\beta$ -farnesene synthase (BFS), which converts farnesyl pyrophosphate (FPP), a linear substrate, into the linear hydrocarbon (E)- $\beta$ -farnesene, an aphid alarm pheromone. In order to probe evolutionary pathways of terpene cyclization, the authors focused on the aa substitutions between BFS and amorpha-4,11-diene synthase (ADS), which produces amorpha-4,11-diene, the bicyclic hydrocarbon precursor of artemisinin, from FPP. Structure-based combinatorial protein engineering (SCOPE) (Dokarry et al. 2012) was employed to construct a library of soluble and biochemically active mutant enzymes with ADS substitutions within 6 Å of the BFS active site. The resulting library contained 93 enzymes with mutations at up to 25 positions with respect to BFS (Salmon et al. 2015). Amino acids at each variable site were restricted to be either in wild type (BFS) or mutant (ADS) state, resulting in a two-letter aa alphabet. Each enzyme in the library, denoted M25, was biochemically characterized using gas chromatography–mass spectrometry (GC-MS) (Garrett et al. 2012), which allowed to determine the spectrum of terpene products, and the malachite green assay (MGA), which was used to measure the total kinetic rate  $k_{\text{cat}}$  (Vardakou et al. 2014). Taken together, these experiments yielded measurements of kinetic rates  $k_{\text{cat},i}$  for 11 terpene products, both cyclic and linear.

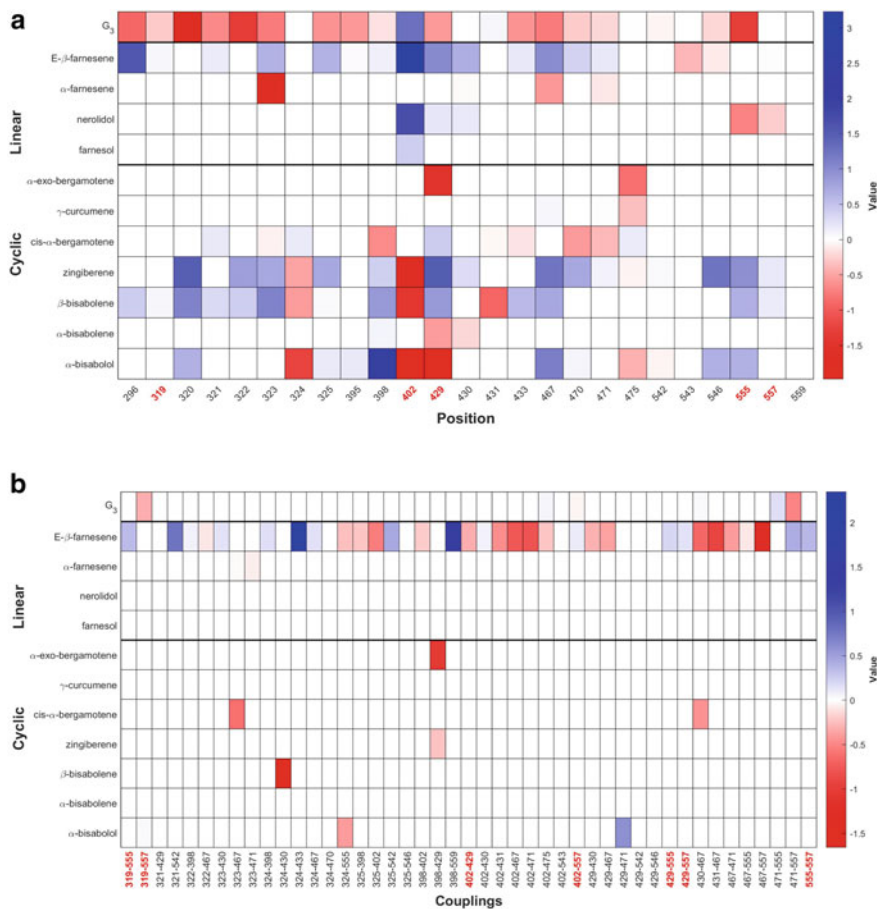
Chief among these products, with detectable levels of catalysis in multiple enzyme variants, was  $\alpha$ -bisabolol, a cyclic terpene alcohol which is the product of a dedicated TPS enzyme in *A. annua* and other *Asteracea* plants. One M25 variant, BOS, that contained five amino acid substitutions with respect to BFS produced especially high levels of  $\alpha$ -bisabolol: 61% of the total output. In order to identify which residues and residue combinations were responsible for the emergence of  $\alpha$ -bisabolol activity, another library, M5, was designed (Ballal et al. 2020). This library consisted of all

combinations of the five amino acid substitutions between BFS and BOS, at positions 319, 402, 429, 555, and 557 in the *A. annua* BFS sequence ( $2^5 = 32$  sequences in total). Similar to the enzymes in the M25 library, each enzyme in the M5 library was assayed using GC-MS and MGA, and the kinetic rates for the 11 terpene products were determined.

The combined M5+M25 library contained  $N = 122$  distinct sequences, including wild-type BFS. Although this library represents only a tiny fraction of the exponentially large number of all possible protein sequences, it was sufficiently large to train a model for Michaelis-Menten free energies  $G_3$  and  $G_{4,i}$  which could describe functional behavior of sequence variants on the two-aa subspace in the vicinity of BFS. Each free energy was represented by the sum of one-body and two-body terms in Eq. (2.2). To reduce the number of fitting parameters, the LASSO constraint was employed and all terms containing one or two wild-type amino acids were set to zero (Morcos et al. 2011). Based on our previous experience with free energy models for protein folding and binding, we expected this approach, which considers Michaelis-Menten free energies rather than kinetic rates as phenotypes, to yield sparse models with just a few two-body terms. Indeed, after fitting the model to the kinetic rate data from the combined library, we found that 12 free energy landscapes (one for  $G_3$  and eleven more for  $G_{4,i}$  are fully described by 113 out of  $12 \times 25 = 300$  possible one-body terms and just 54 out of  $12 \times 138 = 1656$  possible two-body couplings, where 138 is the number of amino acid pairs in the combined dataset for which all four aa combinations are available: (W,W), (W,M), (M,W), and (M,M) (Fig. 2.3). Despite being very sparse, the model is capable of reproducing the observed  $k_{cat,i}$  values with  $r^2 = 0.99$  (Ballal et al. 2020).

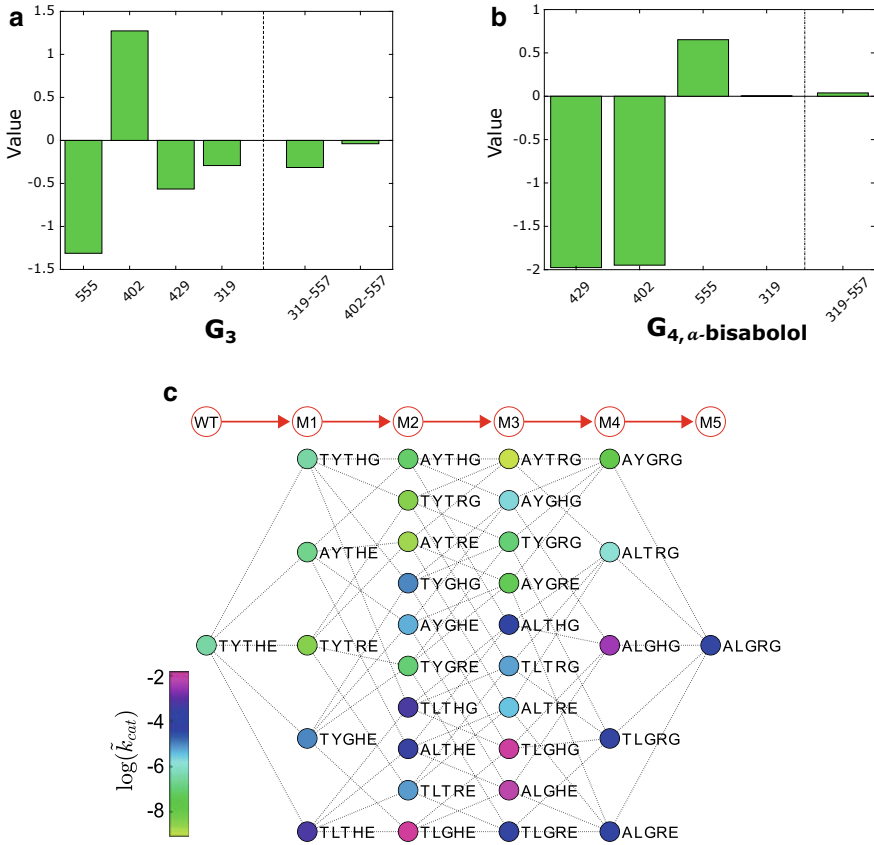
Interestingly, changes in  $G_3$  were predominantly negative, implying that the BFS-to-BOS mutations tend to have adverse effects on the overall values of reaction rates [(Eq. (2.9)]. The only exception to this rule is the Y402L mutation which both increases the total reaction rate and tilts the enzyme specificity toward the cyclic products by lowering the relative reaction rates for linear products (E)- $\beta$ -farnesene (the original BFS product) and nerolidol and by increasing the relative reaction rates for cyclic products zingiberene,  $\beta$ -bisabolene, and  $\alpha$ -bisabolol (Fig. 2.3a). Thus, Y402L plays a role of a “gateway” cyclization-unlocking mutation between enzymes producing linear and cyclic products (Salmon et al. 2015). Other key positions that promote production of cyclic products are 324 and 429. Finally, note that mutations at 10 out of 25 positions result in single-aa terms that suppress (E)- $\beta$ -farnesene production. In addition to single-aa terms, the structure of the free energy landscapes is shaped by 48 two-body couplings which affect  $G_{4,i}$  values and another 6 which correspond to  $G_3$  (Fig. 2.3b). Interestingly, out of the 48 two-aa  $G_{4,i}$  terms, 10 increase reaction rates for cyclic products and only 2 decrease those rates; for linear products, 21 terms contribute to a rate increase and 15 to a rate decrease. Thus, on average, two-body terms tend to promote cyclization.

Focusing now on the free energy landscapes for the M5 library, we observe that for this sequence subset there are only four one-body and two two-body terms that shape the  $G_3$  landscape (Fig. 2.4a), and four one-body and one two-body term that contribute to the  $G_{4,\alpha\text{-bisabolol}}$  landscape (Fig. 2.4b). All other parameters are zero



**Fig. 2.3** One-aa and two-aa contributions to Michaelis-Menten free energies. Fitted values of one-aa (A) and two-aa (B) model parameters for one  $G_3$  and eleven  $G_{4,i}$  landscapes labeled by the product type. Variable positions and pairs of positions in the M5 library is highlighted in red. All free energies are shown in units of  $k_B T$

either by construction or set to zero by the LASSO fit. As a result, the landscape of  $G_3(\sigma) - G_{4,\alpha\text{-bisabolol}}(\sigma)$  values, which determine  $\log(k_{\text{cat},\alpha\text{-bisabolol}}(\sigma))$  up to an additive constant (note that  $\beta = 1$  here, so that all Michaelis-Menten free energies are measured in units of  $k_B T$ ), has a simple, interpretable structure (Fig. 2.4c). Indeed, the structure of the free energy landscape is largely determined by the states of amino acids at positions 402 and 429, with position 555 playing a secondary role. As explained above, the Y402L mutation is key to  $\alpha$ -bisabolol production (compare log kinetic rates for TYTHE and TLTHE in Fig. 2.4c). Another important mutation is T429G (compare, e.g., log kinetic rates for TLTHE and TLGHE in Fig. 2.4c). The TLGHE double mutant favors both  $\alpha$ -bisabolol production and high overall



**Fig. 2.4** Michaelis-Menten free energy landscape. **a** Values of all nonzero one- and two-body parameters in the pairwise expansion of  $G_3$  (first three terms on the right-hand side of Eq. (2.2)) for 32 sequences in the M5 library. The values are sorted from left to right by their absolute magnitude, separately for one- and two-body terms. **b** Same as (a), for  $G_4$  of  $\alpha$ -bisabolol. **c** The logarithm of the  $\alpha$ -bisabolol production rate in the M5 library (Eq. (2.9)). Each node on the landscape is labeled by a string of amino acids at the five variable positions: 319, 402, 429, 555, and 557. Nodes that differ by a single amino acid substitution are connected by a dotted line. The arrows and the circles above the landscape indicate the number of mutations away from the wild-type *A. annua* BFS sequence TYTHE. All sequences are color-coded according to their predicted  $\log(\hat{k}_{cat, \alpha\text{-bisabolol}})$  values, where  $\hat{k}_{cat, \alpha\text{-bisabolol}}(\sigma) = k_{cat, \alpha\text{-bisabolol}}(\sigma) / k_{cat}(\text{WT})$  ( $k_{cat}(\text{WT})$  is the observed total reaction rate of wild-type BFS)

output, and in fact corresponds to the global maximum of the phenotypic landscape in Fig. 2.4c, with no competing local maxima.

We note two additional factors that shape subsequent analysis: (i) All free energy differences that are  $\ll 1$  are not expected to affect the thermodynamic behavior of the ensemble of enzymes inside a cell because such differences are much less than  $k_B T$ , the scale of thermal fluctuations; (ii) Even thermodynamically significant free energy

differences may map onto fitness differences that are much weaker than the competing evolutionary forces of mutation and selection. In other words, the significance of the forces of selection must be judged in the light of evolutionary dynamics of the entire population. For example, as discussed above, a fitness landscape that appears quasi-neutral to a small population may be non-neutral for a large population in which the stochastic forces of genetic drift, which scale inversely with the (effective) population size (Gillespie 2004), are much weaker.

**Fitness dependence on enzymatic function.** The next step in our procedure requires constructing a fitness landscapes as an explicit function of Michaelis-Menten free energies. As with previously discussed models that employ binding and folding free energies as quantitative traits, the functional form of the connection between phenotypes and fitness can be provided by biophysical considerations, with free parameters learned from evolutionary data. For example, under the assumption that all enzyme products can be divided into beneficial and deleterious, one can postulate the fitness function as

$$\mathcal{F}(\sigma) = \sum_{i \in b} \alpha_i n_i(\sigma) - \sum_{i \in d} \beta_i n_i(\sigma), \quad (2.10)$$

where the sums are over the beneficial and deleterious types of products, respectively,  $n_i$  is the number of product molecules of type  $i$  produced per unit time, and  $\alpha_i > 0$ ,  $\beta_i > 0$  are the corresponding fitness gains and losses per product molecule. Note that in general  $\alpha_i = \alpha'_i - \gamma$ ,  $\beta_i = \beta'_i + \gamma$ , where  $\gamma > 0$  is the fitness cost of making or acquiring a single substrate molecule. By construction, benefits outweigh substrate-related costs for all beneficial products. If necessary, Eq. (2.10) can be simplified further by assuming that all fitness gains and losses are product-independent:

$$\mathcal{F}(\sigma) = \alpha \sum_{i \in b} n_i(\sigma) - \beta \sum_{i \in d} n_i(\sigma). \quad (2.11)$$

The model in Eq. (2.10) can also be generalized by assuming that fitness is a nonlinear function of the number of product molecules per unit time:

$$\mathcal{F}(\sigma) = \sum_{i \in b} f_i(n_i(\sigma)) - \sum_{i \in d} g_i(n_i(\sigma)), \quad (2.12)$$

where the activation functions  $f_i$  and  $g_i$  may be sigmoids (if it appears plausible that excessive product molecules simply saturate fitness gains or losses) or Gaussians (if it is more likely that fitness is maximized only if  $n_i$  is close to optimal, and deviations from optimal production rates lead to fitness losses). In the best-case scenario, it should be possible to infer not only the parameters characterizing a given functional form (such as the mean and the variance of the Gaussian), but the functional forms themselves from the observed growth rates of the population into which various mutant enzymes have been introduced.

In contrast to the functions  $f_i$  and  $g_i$  that map enzymatic production rates  $n_i$  onto fitness, the production rates themselves do not require evolutionary information and can be fully expressed in terms of Michaelis-Menten free energies and other biochemical quantities. Indeed, within the Michaelis-Menten framework each  $n_i$  is given by the reaction velocity per enzyme molecule (Nelson 2007):

$$n_i(\sigma) = k_{\text{cat},i}(\sigma) \frac{c}{K_{M,i} + c}, \quad (2.13)$$

where  $K_{M,i}$  is the Michaelis constant of product  $i$  and  $c$  is the substrate concentration, for simplicity assumed to be constant over the time scales of interest (this assumption can be relaxed if substrate concentration data indicates otherwise). In the high substrate-concentration limit ( $c \gg K_{M,i}, \forall i$ ), the reaction velocity reaches its maximum value:  $n_i(\sigma) \simeq k_{\text{cat},i}(\sigma)$ . Thus,  $n_i$  can be immediately expressed through  $G_3$  and  $G_{4,i}$  using Eq. (2.9). In the low substrate-concentration limit,  $n_i(\sigma) \simeq ck_{\text{cat},i}(\sigma)/K_{M,i}(\sigma)$  and the fitness model requires an additional model for the Michaelis constants  $K_{M,i} = (k_{-1} + k_{\text{cat},i})/k_1$ , where  $k_1$  and  $k_{-1}$  can be expressed in terms of MM free energies  $G_1$ ,  $G_2$ , and  $G_3$  (Fig. 2.2) (Bozlee 2007). Note that if  $k_{\text{cat},i} \gg k_{-1}$ , the production rate becomes  $n_i \simeq ck_1$ , such that the overall production rate and, by extension, fitness depend only on the substrate concentration and the height of the  $G_1 - G_2$  free energy barrier.

In summary, just as with the quantitative traits of protein folding and ligand binding, it is useful to employ intermediate molecular phenotypes in assigning fitness values to genomes that harbor enzyme mutant sequences. In this case, the phenotypes are naturally provided by the Michaelis-Menten theory of enzyme kinetics, which yields a set of free energies that describe enzymatic function. These free energies, just like protein folding and ligand-binding free energies, admit a compact description which can be learned from a relatively modest-size library of mutant enzymes whose reaction rates are available experimentally. Then, fitness differences imparted by the mutations in the enzymatic sequence can be described as nonlinear functions of these Michaelis-Menten free energies. There are in fact two types of nonlinearities in the model: one occurs when the free energies are converted into product-type-dependent reaction rates and Michaelis constants, and the other appears when these quantities are translated into fitness. In the approach described above, the former nonlinearity is imposed by biophysical theory while the latter one is to be learned from evolutionary data. There are of course many variations of this procedure, with different forms of fitness functions tested against evolutionary data via machine learning.

## 2.4 Discussion and Conclusion

In this review, we have proposed a two-tiered approach to modeling evolutionary dynamics of populations on fitness landscapes. Instead of focusing on mapping genotypes to fitness values directly, which may easily become daunting due to the



complexity of the fitness landscape, we have demonstrated that, at least in the context of protein evolution, it is beneficial to start by mapping genotypes onto one or several phenotypic landscapes. The usefulness of these “intermediate” phenotypic maps comes from the sparsity of the underlying mathematical models—it appears to be possible to learn very compact descriptions relating phenotypic values to protein sequences, at least for free-energy-based phenotypes. Indeed, the simplest yet informative model of protein folding free energy is a sum of independent contributions of each residue. Such a model can be learned by simply mutating an amino acid in the reference sequence into all other amino acids at every position and measuring the resulting  $\Delta G$  values. Although fairly laborious to collect, such mutational scanning datasets are already available for several proteins, including  $\beta$ -lactamase which imparts ampicillin resistance to *E. coli* (reviewed in Canale et al. (2018)). Such datasets will become more and more common with the advent of more efficient high-throughput molecular biology techniques. As we have shown using evolution of cyclization in terpene synthases as an example, additive phenotypic models can be improved upon by including second-order terms. The key observation is that only a few nonzero pairwise couplings were necessary, such that the free energy models were still nearly additive and retained their compact character.

In the second step of the proposed approach, fitness landscapes are constructed as an explicit function of one or several phenotypes. In some cases, there are compelling biophysical arguments for a specific functional form that relates phenotypic values to fitness, while in others it is necessary to infer the functional form from evolutionary data (that is, fitness measurements for a set of known genotypes). In any event, the key hypothesis is that the resulting fitness model will be much more interpretable compared to a fitness model built directly from genotypes. To probe evolutionary dynamics, fitness landscapes can be explored using either standard computational and analytical tools of population genetics which consider mutation and selection at the sequence level, or first assessing phenotypic effects of sequence mutations and then reconstructing evolutionary dynamics on the fitness landscape as a function of phenotypic changes (Nourmohammad et al. 2013).

Besides studying evolution of proteins and protein complexes (including enzymes), the two-tiered approach can be readily extended to other types of molecular systems. Indeed, the function of many biomolecular machines such as transporters or molecular motors is most naturally described using free energy landscapes (Nelson 2007; Zuckerman 2020). Since all biomolecular free energy landscapes are ultimately based on the same fundamental rules of interatomic interactions, they can presumably be represented using similarly compact, interpretable models which will serve as the basis for constructing fitness landscapes. We note that the two-tiered strategy may be useful well beyond the field of molecular evolution *per se*, with fitness expressed in terms of macroscopic morphological features such as body size and beak shape in Darwin’s finches (Shoval et al. 2012).

Even though evolutionary dynamics appears more interpretable when viewed through the lens of molecular phenotypes or morphological features, it need not be trivial. Indeed, molecular traits such as folding and binding are often correlated: the same aa mutation can contribute to both, for example, if the site in question is located

at the ligand-binding interface. In addition, there are long-range effects which do not require energetic coupling between the two traits but are instead mediated by nonlinearities in the fitness function. For example, a destabilizing mutation away from the ligand-binding interface may push the protein over the stability threshold, making it unfold and lose the ability to interact specifically with ligands (Manhart and Morozov 2015a). Such correlated traits add to the richness of evolutionary trajectories even on low-dimensional fitness landscapes expressed in terms of just a few phenotypes. Furthermore, the dynamics on such continuous landscapes is coarse-grained because it is fundamentally driven by discrete mutations (e.g., single-aa mutations on protein sequence spaces) that cause finite rather than arbitrarily small changes in phenotypic values. This underlying discreteness of moves in the phenotypic space may lead to interesting effects, such as creation of multiple peaks on a fitness landscape defined as a smooth monotonic function of several phenotypic features (Manhart and Morozov 2015a).

Although there are many published studies that allow us to elucidate certain aspects of the genotype→phenotype→fitness paradigm, relatively few can enable us to carry out the two-tiered construction in its entirety. In many cases, this is due to the difficulty of obtaining fitness data, especially in multicellular organisms with longer lifespans. For example, in the terpene synthase study described above, there was sufficient biochemical data to reconstruct phenotypic landscapes on a restricted two-letter alphabet (Salmon et al. 2015; Ballal et al. 2020). However, there is no corresponding set of fitness measurements with mutant protein sequences inserted into their native genomic locations—such measurements would have been exceedingly difficult to obtain for these plant enzymes.

In another study, the authors have created a large-scale library of ~52,000 mutants of the green fluorescent protein from *Aequorea victoria*, a bioluminescent jellyfish found off the west coast of North America (Sarkisyan et al. 2016). Each mutant was characterized by its fluorescence level, which the authors call fitness but which is in fact a phenotype—for obvious reasons, the authors have not provided fitness measurements for these aquatic animals, leaving the relationship between fluorescence and fitness an open question. Interestingly, the authors have examined the relationship between fluorescence and folding free energy using machine learning. This analysis yielded a sigmoid-like function strongly reminiscent of Eq. (2.4). A different large-scale study measured fitness of ~65,000 strains of yeast *Saccharomyces cerevisiae* in a high-temperature environment (Li et al. 2016). Each yeast strain carried a variant of the single-copy tRNA<sub>CCU</sub><sup>Arg</sup> gene in its native genomic location. In this case, the authors have mapped genotypes to fitness directly, without employing intermediate phenotypes. However, they have found that fitness was correlated with the computationally predicted fraction of correctly folded tRNA molecules, indicating that considering fitness explicitly as a nonlinear function of the folding free energy would likely be elucidating.

The three examples above are not comprehensive—rather, they have been chosen to highlight the state-of-the-art in large-scale reconstructions of phenotypic and fitness landscapes. We believe that the next step in this field should focus on synthesiz-

ing information about genotypes, phenotypes, and fitness into a unified evolutionary picture. Some of the key questions to be addressed are: (i) How many phenotypes do we need to describe evolution? (ii) Is it still possible to build tractable evolutionary models when the phenotypes of interest are morphological features rather than molecular properties? (iii) What are the roles of correlations between quantitative traits in different biological systems? Answering these questions leads to exciting future directions in evolutionary biology research.

**Acknowledgements** The authors gratefully acknowledge support from the National Science Foundation (award MCB1920914).

## References

- Aita T, Uchiyama H, Inaoka T, Nakajima M, Kokubo T, Husimi Y (2000) Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. *Biopolymers* 54:64–79
- Ayala FJ, Campbell CA (1974) Frequency-dependent selection. *Annu Rev Ecol Syst* 5:115–138
- Ballal A, Lauredon C, Salmon M, Vardakou M, Cheema J, Defernez M, O'Maille PE, Morozov AV (2020) Sparse epistatic patterns in the evolution of terpene synthases. *Mol Biol Evol* 37:1907–1924
- Berg J, Lässig M (2003) Stochastic evolution of transcription factor binding sites. *Biophysics (Moscow)* 48:S36–S44
- Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42
- Bershtein S, Goldin K, Tawfik DS (2008) Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol* 379:1029–1044
- Bishop CM (2006) *Pattern recognition and machine learning*. Springer, New York, USA
- Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874
- Bloom JD, Lu Z, Chen D, Raval A, Venturelli OS, Arnold FH (2007a) Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol* 5:29
- Bloom JD, Raval A, Wilke CO (2007b) Thermodynamics of neutral protein evolution. *Genetics* 175:255–266
- Bloom JD, Romero PA, Lu Z, Arnold FH (2007c) Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol Direct* 2:17
- Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH (2005) Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA* 102:606–611
- Bozlee BJ (2007) Reformulation of the Michaelis-Menten equation: how enzyme-catalyzed reactions depend on Gibbs energy. *J Chem Ed* 84:106–107
- Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461:515–519
- Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. *Curr Opin Struct Biol* 21:441–446
- Canale AS, Cote-Hammarlof PA, Flynn JM, Bolon DNA (2018) Evolutionary mechanisms studied through protein fitness landscapes. *Curr Op Struc Biol* 48:141–148
- Carneiro M, Hartl DL (2010) Adaptive landscapes and protein evolution. *Proc Natl Acad Sci USA* 107:1747–1751
- Creighton TE (1992) *Proteins: structures and molecular properties*. Freeman and Company, New York

- Crow JF, Kimura M (1965) Evolution in sexual and asexual population. *Am Nat* 99:439–450
- Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York
- DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6:678–687
- Derrida B (1980) Random-energy model: limit of a family of disordered models. *Phys Rev Lett* 45:79–82
- Derrida B (1981) Random-energy model: an exactly solvable model of disordered systems. *Phys Rev B* 24:2613–2626
- Dokarry M, Laurendon C, O’Maille PE (2012) Automating gene library synthesis by structure-based combinatorial protein engineering: examples from plant sesquiterpene synthases. *Meth Enzym* 515:21–42
- Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88:7160–7164
- Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148:1667–1686
- Eshel I, Feldman MW (1970) On the evolutionary effect of recombination. *Theor Pop Biol* 1:88–100
- Ewens W (2004) Mathematical population genetics: I. Theoretical introduction, 2nd edn. Springer, Heidelberg
- Finkelstein AV, Ptitsyn O (2002) Protein physics: a course of lectures. Academic Press, London
- Gardino AK, Villali J, Kivenson A, Lei M, Liu CF, Steindel P, Eisenmesser EZ, Labeikovskiy W, Wolf-Watz M, Clarkson MW, Kern D (2009) Transient non-native hydrogen bonds promote activation of a signaling protein. *Cell* 139:1109–1118
- Garrett SR, Morris RJ, O’Maille PE (2012) Steady-state kinetic characterization of sesquiterpene synthases by gas chromatography-mass spectroscopy. *Meth Enzym* 515:3–19
- Gerland U, Hwa T (2002) On the selection and evolution of regulatory DNA motifs. *J Mol Evol* 55:386–400
- Gillespie J (2004) Population genetics: a concise guide. The Johns Hopkins University Press, Baltimore, USA
- Haldane A, Manhart M, Morozov AV (2014) Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol* 10:e1003683
- Hartl D, Clark A (2007) Principles of population genetics, 4th edn. Sinauer Associates
- Hartman EC, Tullman-Ercek D (2019) Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution. *Curr Op Syst Biol* 14:25–31
- Istomin AY, Gromiha MM, Vorov OK, Jacobs DJ, Livesay DR (2008) New insight into long-range nonadditivity within protein double-mutant cycles. *Proteins* 70:915–924
- Kauffman S (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, New York, USA
- Kauffman SA, Weinberger ED (1989) The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J Theor Biol* 141:211–245
- Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK
- Kingman JFC (1978) A simple model for the balance between selection and mutation. *J Appl Probab* 15:1–12
- Kondrashov AS (1988) Deleterious mutations and the evolution of sexual reproduction. *Nature* 336:435–440
- Landau LD, Lifshitz EM (1980) Statistical physics, Part 1, 3rd edn. Elsevier, Oxford, UK
- Lässig M (2007) From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinf* 8:S7
- Levin BR (1988) Frequency-dependent selection in bacterial populations. *Phil Trans R Soc Lond B* 319:459–472
- Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. *Science* 352:837–840

- Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A (2013) Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol* 30:2645–2653
- Lobkovsky AE, Wolf YI, Koonin EV (2011) Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput Biol* 7:e1002302
- Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, Thomas WK (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 105:9272–9277
- MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge, UK
- Manhart M, Morozov AV (2013) Path-based approach to random walks on networks characterizes how proteins evolve new functions. *Phys Rev Lett* 111:088102
- Manhart M, Morozov AV (2014) Statistical physics of evolutionary trajectories on fitness landscapes. In: Metzler R, Oshanin G, Redner S (eds) First-passage phenomena and their applications. World Scientific, Singapore
- Manhart M, Morozov AV (2015a) Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proc Natl Acad Sci USA* 112:1797–1802
- Manhart M, Morozov AV (2015b) Scaling properties of evolutionary paths in a biophysical model of protein adaptation. *Phys Biol* 12:045001
- Mezard M, Montanari A (2009) Information, physics, and computation. Oxford University Press, Oxford, UK
- Miton CM, Tokuriki N (2016) How mutational epistasis impairs predictability in protein evolution and design. *Prot Sci* 25:1260–1272
- Monod J, Wyman J, Changeux J-P (1965) On the nature of allosteric transitions: a plausible model. *J Mol Biol* 12:88–118
- Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108:E1293–E1301
- Mustonen V, Kinney J, Callan CG, Lässig M (2008) Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA* 105:12376–12381
- Nelson P (2007) Biological physics: Energy, Information, Life. W.H. Freeman and Company, New York, USA
- Nourmohammad A, Held T, Lässig M (2013) Universality and predictability in molecular quantitative genetics. *Curr Op Genet Dev* 23:684–693
- Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* 445:383–386
- Poelwijk FJ, Tanase-Nicola S, Kiviet DJ, Tans SJ (2011) Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol* 272:141–144
- Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10:866–876
- Salmon M, Laurendon C, Vardakou M, Cheema J, Defernez M, Green S, Faraldos JA, O'Maille PE (2015) Emergence of terpene cyclization in *Artemisia annua*. *Nat Comm* 6:6143
- Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O et al (2016) Local fitness landscape of the green fluorescent protein. *Nature* 533:397–401
- Sengupta AM, Djordjevic M, Shraiman BI (2002) Specificity and robustness in transcription control networks. *Proc Natl Acad Sci USA* 99:2072–2077
- Serrano L, Day AG, Fersht AR (1993) Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J Mol Biol* 233: 305–312
- Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wrighers W (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330:341–346

- Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336:1157–1160
- Smith JM (1970) Natural selection and the concept of a protein space. *Nature* 225:563–564
- Stadler PF (1996) Landscapes and their correlation functions. *J Math Chem* 20:1–45
- Starr TN, Thornton JW (2016) Epistasis in protein evolution. *Prot Sci* 25:1204–1218
- Szendro IG, Schenk MF, Franke J, Krug J, de Visser JA (2013) Quantitative analyses of empirical fitness landscapes. *J Stat Mech*, p P01005
- Tholl D (2006) Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Curr Op Plant Biol* 9:297–304
- Tibshirani R (1997) The lasso method for variable selection in the Cox model. *Stat Med* 16:385–395
- Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369:1318–1332
- van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA* 96:9716–9720
- Vardakou M, Salmon M, Faraldos JA, O'Maille PE (2014) Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* 1:187–196
- Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* 9:965–974
- Wakeley J (2005) The limits of theoretical population genetics. *Genetics* 169:1–7
- Weinberger E (1990) Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol Cybern* 63:325–336
- Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312:111–114
- Weinreich DM, Watson RA, Chao L (2005) Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59:1165–1174
- Wells JA (1990) Additivity of mutational effects in proteins. *Biochemistry* 29:8509–8517
- Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* 1: 183–186
- Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8:206–216
- Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford, UK
- Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104:16152–16157
- Zhang XJ, Baase WA, Shoichet BK, Wilson KP, Matthews BW (1995) Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Eng* 8:1017–1022
- Zuckerman DM (2020) Key biology you should have learned in physics class: using ideal-gas mixtures to understand biomolecular machines. *Am J Phys* 88:182–193

# Chapter 3

## A Practical Guide to Orthology Resources



Paul de Boissier and Bianca H. Habermann

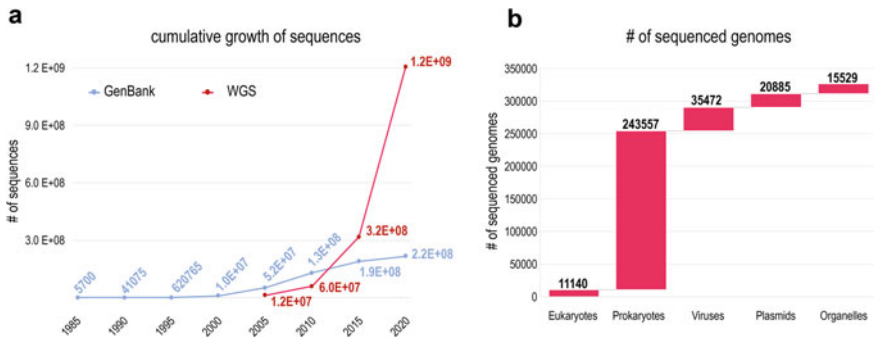
**Abstract** Many resources exist which collect information on orthologous genes and provide this information to the research community. However, the algorithms used to collect orthologs, the type of data available for analysis or download, the range of organisms included, as well as the user-friendliness vary greatly between different orthology databases. In this review, we present a practical guide to the best-known orthology resources: we here briefly discuss their algorithmic details, review their taxonomic coverage and illustrate their user-friendliness. Moreover, we evaluate their capability to detect remotely conserved orthologs and to resolve inparalog relationships in gene families. Moreover, we test them for potential false-positive classification by using a multi-domain protein family with a complex evolutionary history. Finally, we assess the availability and ease of usage of orthology search engines offered by orthology database providers for local usage.

### 3.1 Introduction

With next-generation sequencing (NGS) methods, the number of completely sequenced genomes—and thus the availability of complete proteomes—has increased tremendously (Fig. 3.1). One essential step after genome sequencing is to annotate its gene products and to predict the putative functions of an organism's proteins. The most common method for functional annotation is to infer a protein's function from its related sequences, namely its orthologs from other, already annotated species. The fundamental basis of the concept for transferring functional information across orthologs is the 'ortholog conjecture' or the standard model of phylogenomics (Koonin 2005; Altenhoff et al. 2012). This theory states that orthologs retain the ancestral function, while paralogs tend to rapidly evolve novel functions (Altenhoff et al. 2012). As many organisms will not be studied experimentally, the

---

P. de Boissier · B. H. Habermann (✉)  
Aix-Marseille University, CNRS, IBDM UMR7288, Case 907—Parc Scientifique de Luminy, 163  
Avenue de Luminy, 13009 Marseille, France  
e-mail: [Bianca.HABERMANN@univ-amu.fr](mailto:Bianca.HABERMANN@univ-amu.fr)



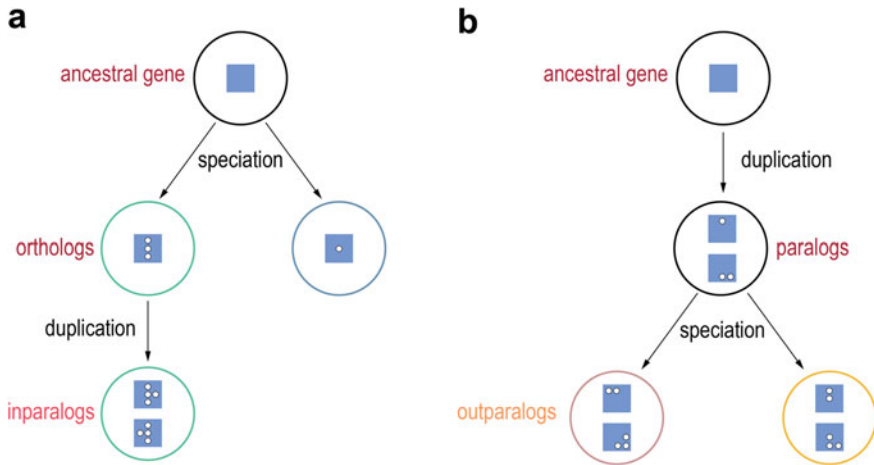
**Fig. 3.1** Growth of sequence databases and completely sequenced genomes in different kingdoms. **a** Cumulative number of sequences found in NCBI GenBank and for whole-genome sequences (WGS) since 1985. **b** Number of sequenced genomes available at NCBI

functional annotation of their genomes relies exclusively on transferring functional knowledge on proteins from other, experimentally studied organisms.

The formal definition of homologous proteins is that they share a common ancestor and thus are homologous on sequence level. Homologs can be divided into different categories depending on their ancestry (Koonin 2005): orthologs, which will be discussed here, result from an event of speciation. Orthologs are typically used to infer gene functions for newly sequenced species. Paralogs are homologous proteins resulting from a gene duplication event. These can be further divided into inparalogs, which result from gene duplication after speciation, and outparalogs, which result from a duplication event before speciation (Fig. 3.2). Finally, xenologs result from horizontal gene transfer. Gene duplication and gene losses, together with horizontal gene transfer, make the distinction of orthologs often difficult, as it is sometimes hard to distinguish, whether a predicted ortholog has arisen from speciation, or from a combination of gene duplication and gene loss events.

There are in principle two types of approaches for identifying orthologs: phylogeny-based methods and methods based on the reciprocal best hit (RBH) theory (Wolf and Koonin 2012). Performing a phylogenetic analysis requires to collect family members, align them, calculate a phylogenetic tree and reconcile the tree for gene gains and losses. Phylogeny-based orthology inference methods tend to be more accurate, as they require a certain amount of manual curation, such as optimizing multiple sequence alignment, and offer a wider choice of parameters, e.g., for tree reconstruction. However, this makes it also harder to compare different phylogeny-based orthology resources (Kriventseva et al. 2008). Furthermore, phylogeny-based methods tend to be computationally expensive. RBH-based methods (which can also be referred to as best reciprocal hit (BRH) or best-best hit (BBH) or genome-specific best hit (BeT)) rely on sequence similarity searches and consider two proteins orthologous if they are each other's best hit in their respective proteomes. They were first introduced with the cluster of orthologous groups (COG) database (Tatusov et al. 2003). RBH-based methods are easy to implement computationally and can be scaled





**Fig. 3.2** Schematic representation of homologous relationships. Orthologs and paralogs are produced, respectively, by speciation and duplication events. Inparalogs and outparalogs are paralogs produced after or before a speciation event. **a** Inparalogs are the result of a direct duplication event. Two inparalogs are therefore more closely related to each other than to any other gene in another organism. **b** Outparalogs on the other hand are the result of a duplication event followed by a speciation event. Colored circles represent species, blue boxes represent a gene and small white circles represent mutations

up to treat hundreds and thousands of genomes. Thus, RBH-based methods made it possible to automatize orthology assignment, and thus, they are at the base of many orthology search engines published to date.

Several tools and databases were created to group and unify genes and proteins based on their evolutionary relationship. These orthology resources are very useful in guiding biologists of different disciplines through the evolutionary history of their proteins of interest. As they use different approaches to collect orthologous proteins, contain different sets of organisms and offer different analysis tools, the information they provide and their user-friendliness differ substantially.

In this chapter, we will focus on orthology resources and aim at helping the reader to find a suitable database for identifying orthologous genes. We will discuss their user-friendliness, their completeness and whether they can resolve problems caused by inparalogs and remote orthologs.

One obstacle in the quest for orthologs is remote orthology. Remote orthologs typically share below 20% sequence identity at protein level; this zone is referred to as the twilight zone of sequence similarity (Walter 1989). Thus, they are difficult to detect with traditional search methods such as BLAST. To discover remote orthologs, more sensitive methods such as profile-based methods have to be used (Steinegger et al. 2019). It is therefore not surprising that remote orthologs are not detected by many orthology search engines. Yet, some search engines do manage to include more remotely conserved orthologs when identifying gene families. In order to probe orthology resources and their underlying algorithms for their ability to detect also

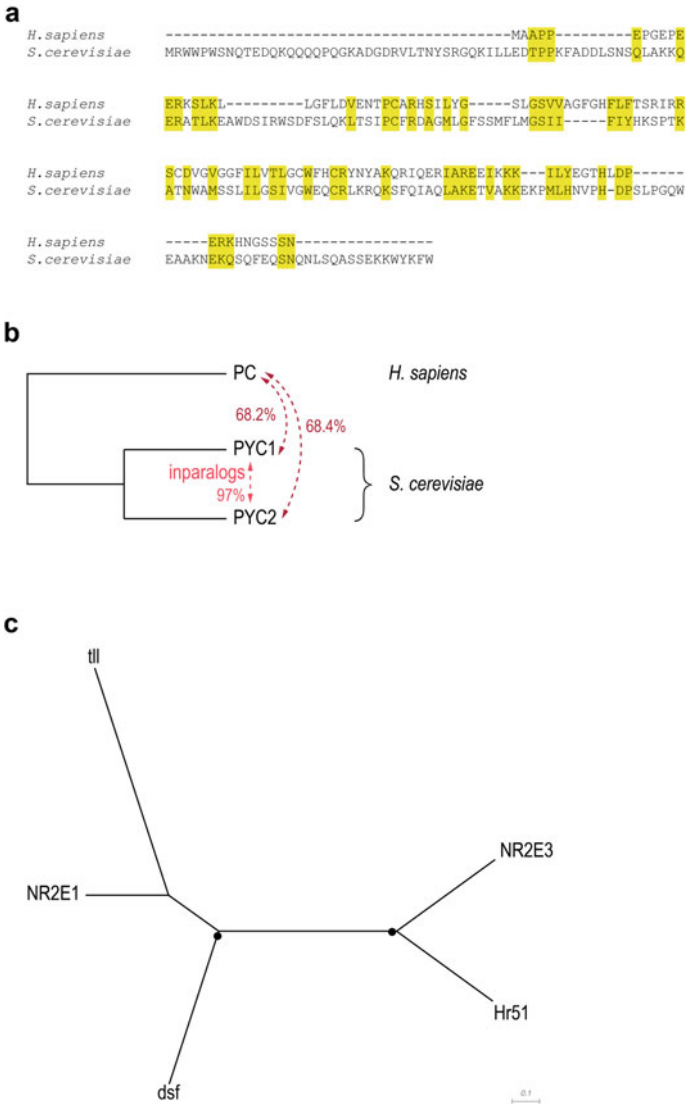
remote orthologs, we have selected the cytochrome C oxidase assembly protein COX20 from *Homo sapiens* (Table 3.1). Human COX20 (aka FAM36A) was found as a remote ortholog of the protein COX20 of the budding yeast *Saccharomyces cerevisiae* (Szklarczyk et al. 2013). Human COX20 is only half the size of its yeast ortholog (118 vs. 205 amino acids (aa), respectively). Submitted to Needle of the EMBOSS software from the EBI (Rice et al. 2000), with a gap open penalty of 10 and an extend penalty of 0.5 and the BLOSUM62 matrix, these two proteins share only 27% sequence similarity and 14% sequence identity, which makes them remote orthologs (Fig. 3.3a). We can therefore use COX20 to estimate the completeness of different databases with respect to proteins with low sequence conservation.

We also wanted to assess orthology resources for their ability to resolve inparalog relationships. Thus, we selected pyruvate carboxylase protein (PC) of *H. sapiens* as our second test case. PC is known to have two inparalogs in *S. cerevisiae* (Pronk et al. 1996), PYC1 and PYC2 (Table 3.1). When submitted to Needle using the same parameters as were used for COX20, human PC has 68.2% sequence similarity to PYC1 from *S. cerevisiae*, and 68.4% sequence similarity to yeast PYC2, respectively. PYC1 and PYC2 are more than 97% similar to each other, which makes them inparalogs, being more similar to each other than to their ortholog(s) in another species (Fig. 3.3b).

Finally, we wanted to investigate putative false-positive assignments. The best candidates for potential false-positive classifications are multi-domain proteins in outparalog relationships. We chose the tailless protein family, which contains a nuclear hormone receptor (NR) domain together with a zinc finger domain. Tailless from *Drosophila melanogaster* has three close paralogs, tailless (tll), dissatisfaction (dsf) and hormone receptor 51 (Hr51). Two human proteins, NR2E1 and

**Table 3.1** Information on the proteins used for testing orthology resources

Organism	Gene Name	Protein sequence ID	mRNA sequence ID	Gene ID	Ensembl ID/locus tag
<i>H. sapiens</i>	COX20 (FAM36A)	NP_001299800	NM_001312871	116228	ENSG00000203667
<i>S. cerevisiae</i>	COX20	NP_010517	NM_001180539	851817	YDR231C
<i>H. sapiens</i>	PC	NP_000911	NM_000920	5091	ENSG00000173599
<i>S. cerevisiae</i>	PYC1	NP_011453	NM_001180927	852818	YGL062W
<i>S. cerevisiae</i>	PYC2	NP_009777	NM_001178566	852519	YBR218C
<i>D. melanogaster</i>	tll	NP_524596	NM_079857	43656	CG1378
<i>D. melanogaster</i>	dsf	NP_477140	NM_057792	33823	CG9019
<i>D. melanogaster</i>	Hr51	NP_611032	NM_137188	36702	CG16801
<i>H. sapiens</i>	NR2E1	NP_001273031	NM_001286102	7101	ENSG00000112333
<i>H. sapiens</i>	NR2E3	NP_057430	NM_016346	10002	ENSG00000278570



**Fig. 3.3** Sequences chosen for testing different orthology resources. **a** Pairwise alignment of COX20 from *H. sapiens* and *S. cerevisiae*. Similar residues are highlighted in yellow. The two proteins share below 30% of sequence similarity and only 14% of sequence identity, which makes them remote orthologs. **b** Tree representation of the relationship between the pyruvate carboxylases. Inparalogs are more similar to each other than their ortholog in *H. sapiens*. **c** Phylogenetic tree of the tailless family, showing the relationship of the three proteins tailless (tll), dissatisfied (dsf) and hormone receptor 51 (Hr51) with the human family members NR2E1 and NR2E3. Proteins were aligned using mafft (Kato and Standley 2013), the tree was calculated using IQ-TREE (Nguyen et al. 2015) with the -s option and 1000 iterations

NR2E3, are equally member of this NR subfamily (Fig. 3.3c). While it is difficult to unequivocally assign orthology in multi-branching families, phylogenetic analysis, in agreement with many orthology resources, assigns NR2E1 as orthologous to tll and NR2E3 as orthologous to Hr51. Needle from EMBOSS reports 51.5% sequence similarity between tll and NR2E1, and only 33.2% sequence similarity between dsf and NR2E1, mostly owing to the fact that the dsf protein contains a long insertion in the center of its sequence and is thus 239 amino acids longer than tll.

## 3.2 OrthoDB

The first database we are discussing is OrthoDB (Kriventseva et al. 2008, 2019). It is referred to as a ‘catalog of orthologs’ and computes orthologs on various levels of the taxonomic hierarchy. OrthoDB relies on the RBH method. It first finds best hits between species using the very fast and sensitive MMseqs2 algorithm (Steinegger and Söding 2017). Clusters of orthologs are then build progressively, with specific e-value cut-offs for triangular RBHs and bidirectional RBHs. Clusters are then further expanded to include inparalogs that are identified as more similar to each other within species than to any protein in another species. OrthoDB has since its introduction embraced the fact that orthologous groups are hierarchical. The procedure to identify orthologs is thus applied at each major radiation of the species taxonomy. As a result, it produces more finely resolved groups of closely related orthologs. Functional annotations are added to each group by summarizing the respective annotations from UniProt, NCBI Gene, InterPro and Gene Ontology. In January 2020, it contained data for 1271 eukaryotes, 6013 prokaryotes (5609 bacteria and 404 archaea) and 6488 viruses for a total of 37 million genes.

To search OrthoDB, the user can perform a simple text search, use identifiers from various databases or a protein sequence. The sequence search is limited to 1000 amino acids, which makes it impossible to search with large protein sequences, such as Titin (~30,000 aa). The advanced search option allows adding specific species to the search, which are presented in a tree-like interface. In the simple text search, the user can specify, if the gene has to be present in all species, in more than 90% or 80%, and if it has to be present in a single copy in all, more than 90% or 80% of species.

When performing a simple text search with the term ‘Cox20,’ OrthoDB returns 246 groups, corresponding to search hits at different taxonomic levels, from the level ‘Eukaryota’ down to sub- and even infraorder levels. The user can thus easily mine orthology relationships at wide taxonomic ranges. The identifier from the NCBI database must be the GeneID (here 116228). The results are more precise as it returns only the COX20 group. The same applies to the sequence-based search. At the Eukaryota level, the COX20 group contains 922 orthologs in 875 species. The group hierarchy is shown in an interactive plot right at the top of the web page (Fig. 3.4a). The annotation of the protein family includes a functional description, Gene Ontology (GO) terms and the evolutionary descriptions, including number of



**Fig. 3.4** COX20 OrthoDB group at Eukaryota level. **a** Interactive group hierarchy split in five different subgroups. **b** Annotations of the orthologous group, including GO terms, InterPro domains and evolutionary information. **c** List of orthologs classified by organisms in a taxonomical, tree-like structure. Taxonomical levels can be displayed or hidden by making use of the arrow. For human COX20, all associated information is shown; identifiers are linked to their respective database of origin

copies per organisms, evolutionary rate and gene architecture (Fig. 3.4b). Orthologs by organisms are listed in the third part of the page, with a link to each protein entry at UniProt and InterPro (Fig. 3.4c). At the bottom of the page, the sibling groups are listed with % overlap and InterPro domains (not shown).

Searching for pyruvate carboxylase (PC) orthologs was done using the GeneID. The search could not be performed using the sequence, as the protein has more than 1000 aa. A text-based search with the gene name 'PC' is less accurate, as many groups contain these two consecutive characters. Search results for the PC GeneID (5091) revealed that OrthoDB is able to resolve inparalog relationships, as the database returned PC for *H. sapiens* and PYC1 and PYC2 for *S. cerevisiae*. Only the naming of the group in OrthoDB is confusing, as it is referred to as *biotin carboxylase, C-terminal* in the Eukaryota group, and to *pyruvate carboxylase* starting from Metazoa. In total, this group contains 3888 genes found in 1188 species.

We searched for tailless with its GeneID from NCBI (43656), which returned in Eukaryota the nuclear hormone receptor, ligand binding domain group (870262at2759) with 1305 genes in 444 organisms, including the three *Drosophila* (tll, dsf and Hr51) and two human proteins (NR2E1, NR2E3). More detailed information on the relationship between these five proteins is not returned.

In conclusion, we find that this tool provides extensive information and accurate orthology assignments. It is fast and user-friendly. It succeeded in finding the remote ortholog of COX20 and included the inparalogs from the pyruvate carboxylase family. OrthoDB does not provide detailed information on the phylogenetic relationship of the five proteins of the tailless family, however correctly and exclusively identified them as being part of the same group. Search options are manifold, though the easiest and most precise results are returned when searching either with the NCBI GeneID, or the sequence. OrthoDB is available at: <https://www.orthodb.org/>.

### 3.3 HomoloGene

HomoloGene (NCBI Resource Coordinators 2016, 2018) is a tool developed by the NCBI to detect paralogs as well as orthologs. It contains 21 completely sequenced eukaryotic genomes and profits from the entire information content of the NCBI databases, including in-depth information provided for annotated genes at NCBI (synonyms, gene description, genomic location, isoforms, Gene Ontology (GO) information, interaction partners or literature). HomoloGene uses sequence similarity based on BLASTp (Altschul et al. 1997) comparisons to match sequences into groups using a species tree. More closely related sequences are matched first, followed by more distantly related ones. The algorithm for sequence matching is heuristic and performs bipartite matching, an algorithm derived from graph theory (Bondy and Murty 1976). The matching procedure employed by HomoloGene optimizes the global, rather than the local score of the bipartite graph. For each match, a statistical significance is calculated. Protein alignments are mapped back to their

respective DNA sequences to obtain Ka/Ks ratios: the ratio of the number of substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) over a given time frame. These Ka/Ks ratios are used to filter out sequences that have potentially been incorrectly grouped. Inparalogs are determined by identifying sequences that are more closely related to a sequence within one organism than to a sequence in another organism.<sup>1</sup>

For searching HomoloGene, all basic and advanced search options are available, such as searching different fields, AND/OR options, drop-down lists, etc. We used the basic search for COX20, which returned two HomoloGene groups: an ‘unnamed protein’ group in *Saccharomycetaceae*, which corresponds to COX20 in *S. cerevisiae* and closely related fungi, and the ‘Cox2 chaperone homolog (*S. cerevisiae*)’ group conserved in *Euteleostomi*, corresponding to COX20 conserved from human to zebrafish (Fig. 3.5a). Protein identifiers, species and gene names are shown. Additionally, proteins are represented graphically with their conserved domains. Identifiers are linked to the gene entry and the Refseq protein entry, respectively; the graphical protein representation is linked to the conserved domain database (CDD, (Lu et al. 2020)) entry of the respective domain(s). Users can view and download the multiple sequence alignment of all family members, which can be useful for further analysis; BLASTp-based pairwise alignments of orthologous proteins (Fig. 3.5b, c) can also be downloaded for further analysis or processing. Individual alignment scores for all pairwise comparisons are accessible from the link ‘Show Pairwise Alignment Scores.’ Finally, there is an exhaustive list of articles linked to proteins from the orthologous group.

The search for inparalogs revealed that HomoloGene groups all paralogs in the same orthology group. When searching for pyruvate carboxylase, the two different proteins of *S. cerevisiae* were part of the same group (Fig. 3.5d). PC has several domains, which are shown in different colors.

HomoloGene does not provide information on the entire tailless family, but rather classifies orthologous pairs. Interestingly, and opposing to phylogenetic analysis and other orthology resources, HomoloGene assigns dsf as orthologous to NR2E1. Tll is classified as conserved in Diptera only, and Hr51 is assigned as the ortholog of NR2E3. A better Ka/Ks ratio could be causative for defining dsf as the ortholog of NR2E1, as there is locally a higher number of identical amino acids in the ZnF and NHR domains (Fig. 3.5e).

In conclusion, the main advantage of HomoloGene is its integration in the NCBI database resources and the high confidence on orthologs inherent in its build procedure. Furthermore, inparalogs are identified and indicated as both being part of one HomoloGene group. The ambiguous tailless family is resolved differently than in other orthology databases: dsf, instead of tll, is considered the ortholog of NR2E1. Tll itself is classified as arthropod-specific, and Hr51 is considered orthologous to NR2E3. Moreover, HomoloGene does not provide an overview of the entire tailless

---

<sup>1</sup>It should be noted at this point that information on the HomoloGene algorithm could not be found even in the pages of the NCBI help desk. Information on the build procedure of HomoloGene is therefore taken from Wikipedia (<https://en.wikipedia.org/wiki/HomoloGene>).

**a**

**HomoloGene:11956. Gene conserved in Euteleostomi**

**Genes**

Genes identified as putative homologs of one another during the construction of HomoloGene.

- COX20, *H.sapiens*
- COX20 cytochrome C oxidase assembly factor
- COX20, *Ptilinodytes*
- COX20 Cox2 chaperone homolog (S. cerevisiae)
- COX20, *M.mullata*
- COX20 Cox2 chaperone homolog (S. cerevisiae)
- COX20, *C.lupus*
- COX20 Cox2 chaperone homolog (S. cerevisiae)
- COX20, *B.taurus*
- COX20 Cox2 chaperone homolog (S. cerevisiae)
- COX20, *M.musculus*
- COX20 Cox2 chaperone
- COX20, *R.norvegicus*
- COX20 cytochrome C oxidase assembly factor
- COX20, *G.gallus*
- COX20 Cox2 chaperone homolog (S. cerevisiae)
- LOC100497666, *X.tropicalis*
- cytochrome c oxidase protein 20 homolog
- zgc:92568, *D.rosio*
- zgc:92598

**Proteins**

Proteins used in sequence comparisons and their conserved domain architectures.

- NP\_932342.1
- 118 aa
- XP\_003949782.1
- 130 aa
- XP\_001089471.1
- 118 aa
- XP\_537221.1
- 163 aa
- NP\_001036751.1
- 112 aa
- NP\_079787.1
- 117 aa
- NP\_001099446.1
- 117 aa
- NP\_001264593.1
- 117 aa
- XP\_002942240.1
- 117 aa
- NP\_001002712.1
- 101 aa

**b**

**Protein Alignments**

Protein multiple alignment, pairwise similarity scores and evolutionary distances.

Show Multiple Alignment

Show Pairwise Alignment Scores

Pairwise alignments generated using BLAST

Regenerate Alignments

NP\_932342.1 (H.sapiens)

XP\_003949782.1 (P.troglodytes)

**Publications**

Articles associated with genes and sequences of this homology group.

A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia. Szklarzak R, et al. Hum Mol Genet 22, 656-67 (2013).

**c**

NP_932342.1	7	-PGEFER-----KSLFLGLDLOVENTCANRSLLYGLSDVV	43
XP_003949782.1	7	-PGEFERKACVLSLHFNKVLGLDLOVENTCANRSLLYGLSDVV	55
XP_001089471.1	7	-PGEFER-----KAFLLGLDLOVENTCANRSLLYGLSDVV	43
XP_537221.1	7	-PGEFER-----LFFLLGLDLOVENTCANRSLLYGLSDVV	87
NP_001036751.1	7	-----PFFLLGLDLOVENTCANRSLLYGLSDVV	35
NP_079787.1	7	-PFEETR-----KFFLLGLDLOVENTCANRSLLYGLSDVV	42
NP_001099446.1	7	-PFEER-----KFFLLGLDLOVENTCANRSLLYGLSDVV	42
NP_001264593.1	4	-EGDEFE-----KFFLLGLDLOVENTCANRSLLYGLSDVV	40
XP_002942240.1	1	-----KFFLLGLDLOVENTCANRSLLYGLSDVV	28
NP_001002712.1	5	-EGEYVR-----KFFLLGLDLOVENTCANRSLLYGLSDVV	41

NP_932342.1	44	AGDFLFLTRIRRACDVGWGFILTLQWCRKRYFNANQIGERIANE	93
XP_003949782.1	56	AGDFLFLTRIRRACDVGWGFILTLQWCRKRYFNANQIGERIANE	105
XP_001089471.1	44	AGDFLFLTRIRRACDVGWGFILTLQWCRKRYFNANQIGERIANE	93
XP_537221.1	88	AGDFLFLTRIRRACDVGWGFILTLQWCRKRYFNANQIGERIANE	137
NP_001036751.1	34	AGDFLFLTRIRRACDVGWGFIVTLQWCRKRYFNANQIGERIANE	85
NP_079787.1	43	TGLGFLVTRIRRACDVGWGFILTLQWCRKRYFNANQIGERIANE	93
NP_001099446.1	43	TGLGFLVTRIRRACDVGWGFILTLQWCRKRYFNANQIGERIANE	92
NP_001264593.1	41	YGLGFLVTRIRRACDVGWGFITLQWCRKRYFNANQIGERIANE	90
NP_001002712.1	29	AGLFLVTRIRRACDVGWGFITLQWCRKRYFNANQIGERIANE	78
XP_002942240.1	42	LGLGFLVTRIRRACDVGWGFITLQWCRKRYFNANQIGERIANE	91

NP_932342.1	94	EIKKKILYEGTHLPERKKNRGGSSD--	118
XP_003949782.1	104	EIKKKILYEGTHLPERKKNRGGSSD--	130
XP_001089471.1	94	EIKKKILYEGTHLPERKKNRGGSSD--	118
XP_537221.1	138	GIKKKILYEGTHLPERKKNRGGSSD--	163
NP_001036751.1	86	GIKKKILYEGTHLPERKKNRGGSSD--	112
NP_079787.1	93	GIKKKILYEGTHLPERKKNRGGSSD--	117
NP_001099446.1	93	GIKKKILYEGTHLPERKKNRGGSSD--	117
NP_001264593.1	91	GKNNKLPFGSDFPKKPKKNGRNNNS	117
NP_001002712.1	79	GLNNKLYEGTHLPERKKNRGGSSD--	101
XP_002942240.1	92	GKNNKLPFGSDFPKKPKKNGRNNNS	117

**d**

**HomoloGene:5422. Gene conserved in Opisthokonta**

**Genes**

Genes identified as putative homologs of one another during the construction of HomoloGene.

- PC, *H.sapiens*
- pyruvate carboxylase
- >>>
- PYC2, *S.cerevisiae*
- PYC2
- PYC1, *S.cerevisiae*
- PYC1

**Proteins**

Proteins used in sequence comparisons and their conserved domain architectures.

- NP\_071504.2
- 1178 aa
- NP\_009777.1
- 1180 aa
- NP\_011453.1
- 1178 aa

**e**

**tailless [Drosophila melanogaster]**

ei: 3e+05 | 56: 177(431613) | post: 243(433581) | gpaai: 64(433163)

Query 47	RLL-DIKPVCGDSSGSRDITVACDGGDFPRRIRSRRTVYCRGQDQVPTKRM	105
Hit 1	RLL-VKRCV-D-SRRSRI-VKCGD-GFPRRIRSRRTVYCRGQDQVPTKRM	105
Subject 21	RLLVYKPVCGDSSGSRDITVACDGGDFPRRIRSRRTVYCRGQDQVPTKRM	86

**dissatisfaction, isoform B [Drosophila melanogaster]**

ei: 3e+05 | 56: 172(38173) | post: 80(38181) | gpaai: 2(38121)

Query 45	GRRLLDKPVCGDSSGSRDITVACDGGDFPRRIRSRRTVYCRGQDQVPTKRM	102
Hit 5	GRRLLDKPVCGDSSGSRDITVACDGGDFPRRIRSRRTVYCRGQDQVPTKRM	84



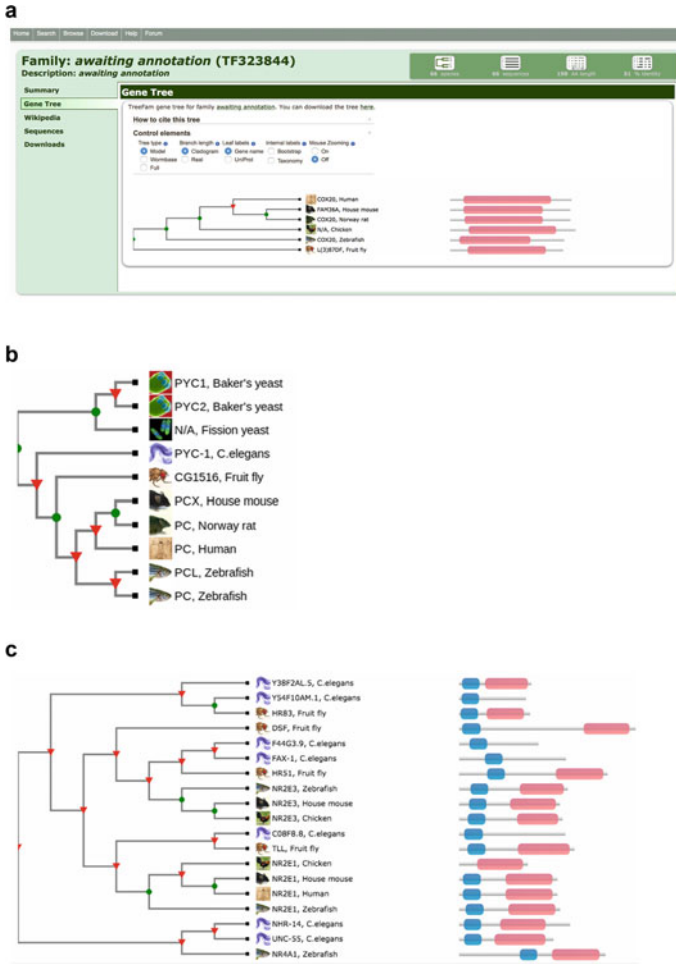
◀**Fig. 3.5** Results of the HomoloGene database. **a** COX20 is conserved in Euteleostomi, which includes *H. sapiens*. Protein domains and sequence lengths are displayed. Links are provided to the gene and the protein entries of the NCBI database. **b** HomoloGene allows to see alignments and launch a pairwise BLAST, or to retrieve pairwise alignment scores. A list of publications linked to the family members is given. **c** Multiple sequence alignment of the protein family. **d** Pyruvate carboxylase results. Diverse domains are colored differently and link to their respective CDD entry. The results page was restricted to show only the results for *H. sapiens* and *S. cerevisiae*. **e** BLASTp alignment of NR2E1 with tailless (tll) and dissatisfaction (dsf). Dsf shows a higher number of identical amino acids in the local alignment, while being the second-best hit found in *D. melanogaster*, when searched with human NR2E1

family, but rather separates tll, dsf and Hr51 into three different families. Disadvantages include the low number of organisms (21 species versus 13722, e.g., in OrthoDB) and the disability to identify remote orthologs. HomoloGene has not been updated since 2014, according to information provided at ‘HomoloGene statistics.’ HomoloGene is available at: <https://www.ncbi.nlm.nih.gov/homologene/>.

### 3.4 TreeFam

TreeFam (Ruan et al. 2008; Schreiber et al. 2014) is one of the bioinformatic tools developed at the EMBL-EBI. The last release (v9, from March 2013) contains 15,736 gene families from 109 species. TreeFam v9 has adopted the Ensembl Compara pipeline to assemble ortholog families, which performs all-against-all BLASTp searches and subsequent clustering of ortholog families. The clustering procedure in TreeFam however uses a hidden Markov model (HMM)-based approach to cluster TreeFam families, which allows the database to have more stable ortholog families with new releases (Schreiber et al. 2014), as new sequences can be added to the existing HMM families. Multiple sequence alignments are created with either MAFFT (Katoh and Standley 2013) or M-Coffee (Wallace et al. 2006) and refined by removing non-conserved positions. TreeBest is used to construct gene trees (<https://treesoft.sourceforge.net/treebest.shtml>). Several trees (based on amino acid as well as back-translated nucleic acid alignments) are constructed, and a consensus tree is calculated using a species tree as a reference. Gene losses and duplications are calculated using the Duplication/Loss Inference algorithm (Li et al. 2006) and by reconciling the tree with the NCBI taxonomy tree (Federhen 2012).

TreeFam can be browsed with a gene name, or searched with a protein sequence. By clicking on the TreeFam family, first, a summary page is invoked, which gives the user information on the general conservation of the query in a species tree. The Gene Tree tab (Fig. 3.6a, found on the right-hand side of the page) displays the gene tree as a model, which can be expanded (by clicking on ‘Full’); the tree can be annotated by adding information on branch length, bootstrap values, labels for taxonomy, etc. The protein nodes are linked to the gene entries of the Ensembl database (Cunningham et al. 2019). Next to the tree are graphical representations of



**Fig. 3.6** TreeFam search showing the model trees for COX20 and PC. **a** Model tree of the COX20 family in TreeFam. Different tabs allow to access different information, such as the sequences and the Wikipedia page. A download page is available to retrieve information associated with the protein family. **b** Reduced phylogenetic tree of the PC family, indicating gene duplication (red arrows) and speciation (green dots) events. **c** TreeFam tree of the tailless family. Hr83, nhr-14, unc-55 as well as several other *C. elegans* proteins are false-positive members of this group

the proteins with identified conserved domains. These link to the respective entry of the conserved domain in the Pfam database (El-Gebali et al. 2019). Wikipedia links to the Wikipedia entry of the gene; the sequences in the tree can be assessed at the ‘Sequences’ link; finally, the alignment in FASTA format, the HMM, as well as the tree in Newick format can be downloaded from the ‘Download’ link. Summary of the gene family is displayed at the top of the entry (listing number of species, sequences, alignment length and % overall identity). When searching with a protein sequence,

the sequence will be grouped to its presumable family by similarity and added to the family tree. The user can choose between two phylogenetic methods: parsimony which is less accurate but faster and maximum likelihood which is slower but more accurate. The sequence is added at the correct position in the tree, however as a separate—duplicated—entry. For instance, when searching with the COX20 protein sequence from human, there will be two identical human nodes in the tree.

COX20 of *S. cerevisiae* is not found in the pre-built tree, though when searching with the budding yeast COX20 protein sequence, the correct family is identified and the sequence is added to the tree.

When searching for the PC family, the two inparalogs in *S. cerevisiae* are correctly placed in the family tree of pyruvate carboxylase. Moreover, the Gene Tree tab allows the user to see, in a small synthetic view with model organisms, where the duplication and speciation events occurred. The red triangle and the green point, respectively, represent the duplication and the speciation events (Fig. 3.6b).

The tailless family contains next to tll, dsf and Hr51 also the protein Hr83 from *Drosophila* (Fig. 3.6c). This protein is not classified as part of the tll family in other databases.

To summarize, this database is quite comprehensive, providing visual display of the family tree, allowing download of underlying alignments and trees and providing functional annotation from Wikipedia. The tree is interactive and can be labeled. Genes are linked to the Ensembl resource, providing rich information on individual genes/proteins. A novel sequence can be added to the family tree. It is able to distinguish inparalogs, in fact indicating speciation and duplication events in the tree. Moreover, tree-based methods are thought to be more powerful in inferring orthology than simple RBH-based approaches (Koonin 2005; Brown and Sjölander 2006). It however did not include *S. cerevisiae* COX20 in the pre-calculated COX20 tree. TreeFam also fails to correctly distinguish orthology relationships of the tailless family, as it assigns false-positive ZnF and NHR-domain containing proteins to this family. Another disadvantage is the lack of recent updates of the database. TreeFam is available at: <https://www.treefam.org/>.

### 3.5 HCOP

HCOP (HUGO Gene Nomenclature Committee (HGNC) Comparison of Orthology Predictions (Eyre et al. 2007)) is a database of 19 species, including *H. sapiens* and *S. cerevisiae*, which allows identification of pairs of predicted orthologs. It combines orthology data from 14 different databases of orthologs, including OrthoDB, OrthoMCL, HomoloGene, OMA, TreeFam, Panther, InParanoid, EggNOG and others. Thus, it integrates orthology data derived from many different search strategies (Fig. 3.7 a). Orthologous pairs from these sources are consolidated into a non-redundant list of orthologs and HCOP provides the associated list of databases that support each assignment. HCOP is human-centric as orthologs of a human protein

**a**

**Search for orthologs between**

Human

**and**

Chimp     Macaque     Mouse     Rat     Dog     Cat  
 Horse     Cow     Opossum     Pig     Platypus     Chicken  
 Anole lizard     Xenopus     Zebrafish     C.elegans     Fruitfly     S.cerevisiae  
 S.pombe

Deselect all

**include orthologs from**

EggNOG     Ensembl     HGNC     HomoloGene     Inparanoid     OMA  
 OrthoDB     OrthoMCL     NCBI     Panther     PomBase     PhyloMeDB  
 Treefam     ZFIN

Deselect all

**where**

Approved symbol(s)

COX20

either enter identifier(s)    or upload file

Browse... No file selected.

Submit

**b**

<b>Human</b>	<p><b>Approved symbol</b> COX20</p> <p><b>Approved name</b> cytochrome c oxidase assembly factor COX20</p> <p><b>Locus type</b> gene with protein product</p> <p><b>Chromosomal location</b> 1q44</p> <p><b>Gene resources</b> <a href="#">HGNC:26970</a> <a href="#">ENSG00000203667</a> <a href="#">116228</a></p>	
<b>Drosophila</b>	<p><b>Gene symbol</b> <a href="#">l(3)87Df</a></p> <p><b>Gene name</b> lethal (3) 87Df</p> <p><b>Locus type</b> protein_coding</p> <p><b>Chromosomal location</b> 3R</p> <p><b>Gene resources</b> <a href="#">FBgn0002354</a> <a href="#">FBgn0002354</a> <a href="#">49762</a></p>	<p><b>Assertion derived from:</b></p>
<b>S.cerevisiae</b>	<p><b>Approved symbol</b> COX20</p> <p><b>Approved name</b> Cytochrome c OXidase</p> <p><b>Locus type</b> protein_coding</p> <p><b>Chromosomal location</b> IV</p> <p><b>Gene resources</b> <a href="#">S000002639</a> <a href="#">YDR231C</a> <a href="#">851817</a></p>	<p><b>Assertion derived from:</b></p>

**Fig. 3.7** HCOP database search page and results. **a** Input options for a HCOP search, presenting all the species and databases from which the information can be extracted. **b** Results of HCOP for COX20 for *Drosophila melanogaster* and *S. cerevisiae*. The results provided are derived from five different databases for the fruit fly; the orthologous budding yeast COX20 protein could only be retrieved from the PANTHER resource

can be found in other species but searching for a protein of another species only returns orthologs in *H. sapiens*.

The search can be done by identifiers of Ensembl, NCBI or HGNC, approved gene symbols or a file containing a list of identifiers. Wildcards can be used: ‘\_’ to substitute a single character and ‘\*’ or ‘%’ for zero, one or several characters. We used the approved symbol COX20. The output is a list of orthologs in the different species which can be saved as a text file. The results give the chromosomal location, and specific identifiers link the ortholog to the database it is found in. The support from

the different orthology resources is furthermore shown for each identified ortholog. Budding yeast COX20 is found as an ortholog and supported by the PANTHER database (as indicated by the PANTHER symbol in the search results). Searching with *S. cerevisiae* COX20 only retrieves the human ortholog (Fig. 3.7b). Searching PC using its official gene name returns all orthologs and inparalogs, including PYC1 and PYC2 from *S. cerevisiae*.

When searching for tll in HCOP, the two human proteins NR2E1 and NR2E3 are found, suggesting that HCOP groups the entire tailless family. When searching for orthologs of human NR2E1, tll, dsf and Hr51 are identified in fruit fly.

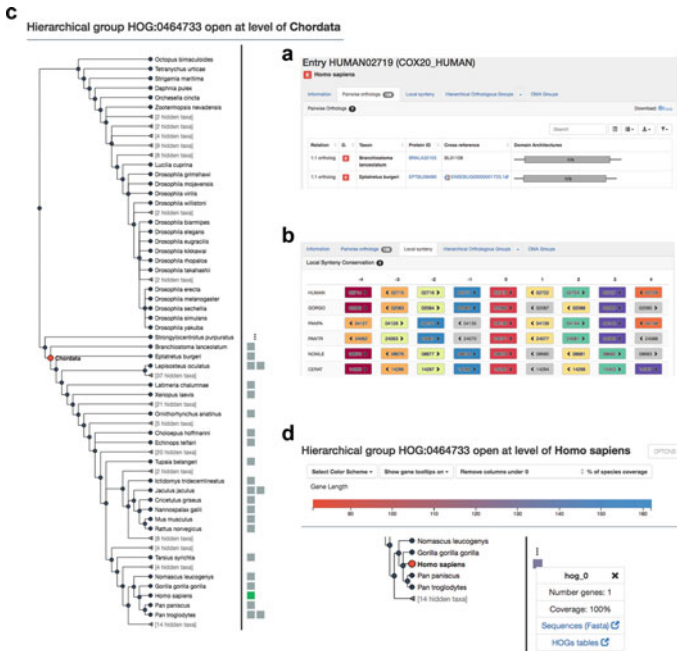
In summary, the HCOP database is a very useful resource as it offers cross-references between different orthology databases. As it relies on data generated by many different search algorithms, it is able to find remote orthologs and includes inparalogs in orthology groups. HCOP limits the tailless family in human and *Drosophila* to the core members. Furthermore, all data can be downloaded in tabular format for local usage. The main disadvantage is that there are only few species included; moreover, it is human-centric and only returns all orthologs when searching with the human sequence. HCOP is available at: <https://www.genenames.org/tools/hcop/>.

## 3.6 OMA

Orthologous Matrix (OMA) was developed at and is hosted by the Swiss Institute of Bioinformatics (SIB) (Altenhoff et al. 2015, 2018). To infer orthologs, it first computes all-against-all Smith-Waterman alignments, saving only candidate pairs with sufficient score and overlap. In the next step, evolutionary distances are used to identify closest homologs, thus defining orthologs based on the reciprocal best hit hypothesis, however considering potential gene losses. Identified orthologs are finally clustered into OMA groups (i.e., most closely related genes between each two species and thus containing only orthologs), which tend to be very specific, as well as hierarchical orthologous groups (HOGs, which are hierarchical groups of all genes that descended from a single common ancestor and thus contain (in)paralogs). A detailed primer of the OMA database and search algorithm is given in a recent review by Zahn-Zabal et al. (2020). OMA is actively maintained (as of 2020) and contains 2288 species (1674 bacteria, 152 archaea and 462 eukaryotes (fungi, animals and plants)). Model organisms are updated at each release, and other genomes are updated at each important re-annotation or added based on user requests. OMA provides domain annotations and synteny data for each gene; moreover, Gene Ontology (GO) terms are inferred for each cluster.

OMA can be searched with either one recognized identifier, an amino acid sequence, an OMA group or by performing a simple text search. Searching with a sequence either performs an ‘exact’ search, returning only hits that match the input sequence exactly, or an ‘approximate’ search, where a few mismatches are allowed. The approximate search is not comparable to BLAST, but rather used for

retrieval of near-identical entries in the database. An approximate search with the protein sequence of human Cox20 returned 127 entries of high homology, representing COX20 1:1 orthologs. When searching with the term COX20 in a free text search, a list of all COX20 proteins is returned. The user can choose to either see the protein record or go directly to identified orthologs. *H. sapiens* COX20 has 139 one-to-one orthologs. A tabs menu lets the users switch from tabular listing of one-to-one (1:1) and one-to-many (m:1) orthologs (Fig. 3.8a), to information on the protein, to local synteny (Fig. 3.8b), to OMA group (downloadable in FASTA format) and HOGs (Fig. 3.8c). By invoking the ‘OPTIONS’ drop-down menu, boxes next to genes in the tree can be colored according to the gene length or the % CG content. Boxes are also interactive, linking to the HOG table of the gene, as well as the sequence (Fig. 3.8d). An alignment can be created, visualized, filtered and downloaded for each OMA group (Fig. 3.8e). Furthermore, a fingerprint is created for each group, representing the most conserved region of its members (Fig. 3.8e, top). For each



**Fig. 3.8** OMA orthology database results pages. **a** Tabular listing of 1:1 orthologs for the human COX20 protein. **b** The local synteny (four genes upward and downward) is displayed by small colored boxes with the direction of transcription. **c** The hierarchical orthologous groups represent the orthologs as a tree with the number of copies per species. The Chordata node is displayed in red. **d** The tree can be decorated with information on the gene length or the GC content (here shown for human COX20). **e** Multiple sequence alignment of the OMA group. A logo is extracted from it, in which the size of the letter for an amino acid is representative for its conservation. **f** OMA groups in which the COX20 proteins from different species are included. **g** GO terms associated with each ortholog

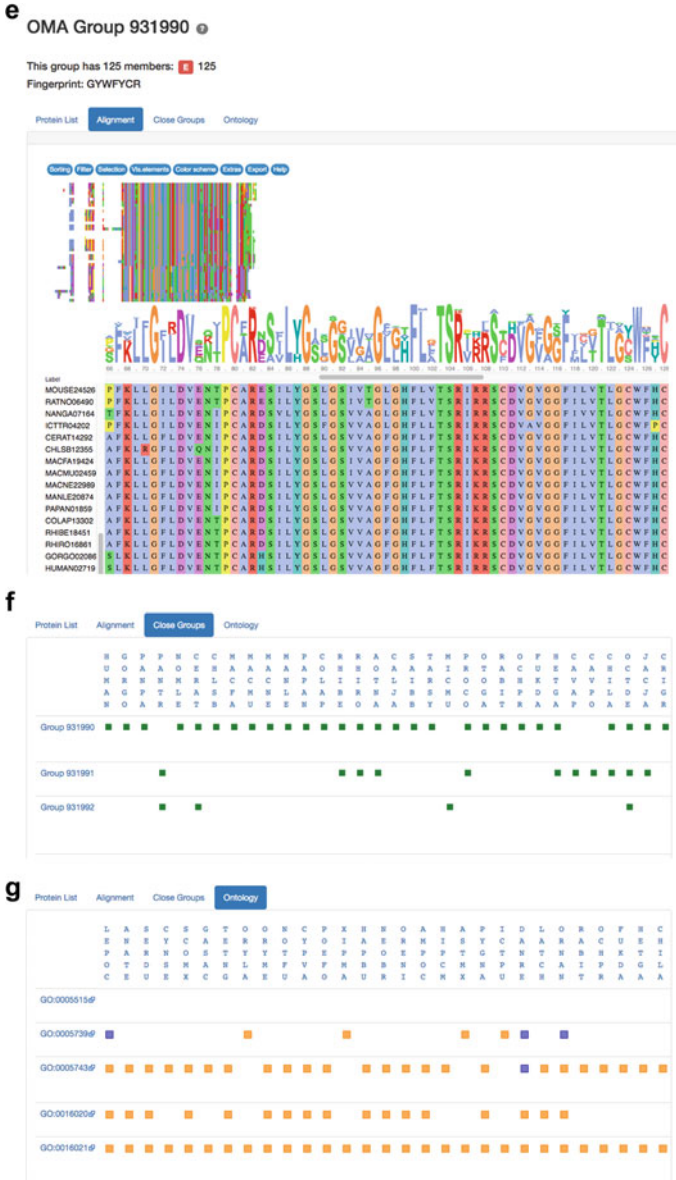


Fig. 3.8 (continued)

species, close OMA groups (Fig. 3.8f) and gene ontologies (Fig. 3.8g) are listed in an easily readable tab format.

OMA was not able to find the remote ortholog from *S. cerevisiae* and the COX20 OMA and HOGs group were confined to metazoans. PC, though present in the database, is not grouped in any OMA or HOG cluster. The two inparalogs PYC1 and PYC2 have nearly 800 1:1 orthologs; they are part of an OMA group of 67 members from bacteria and eukaryotes. The two inparalogs are listed in this group as ‘close paralogs.’

The tailless family is split in separate orthologous groups in OMA. Tll itself is grouped with human NR2E1, and Hr51 is found in the same OMA group as human NR2E3. Dsf is classified as an arthropod-specific protein.

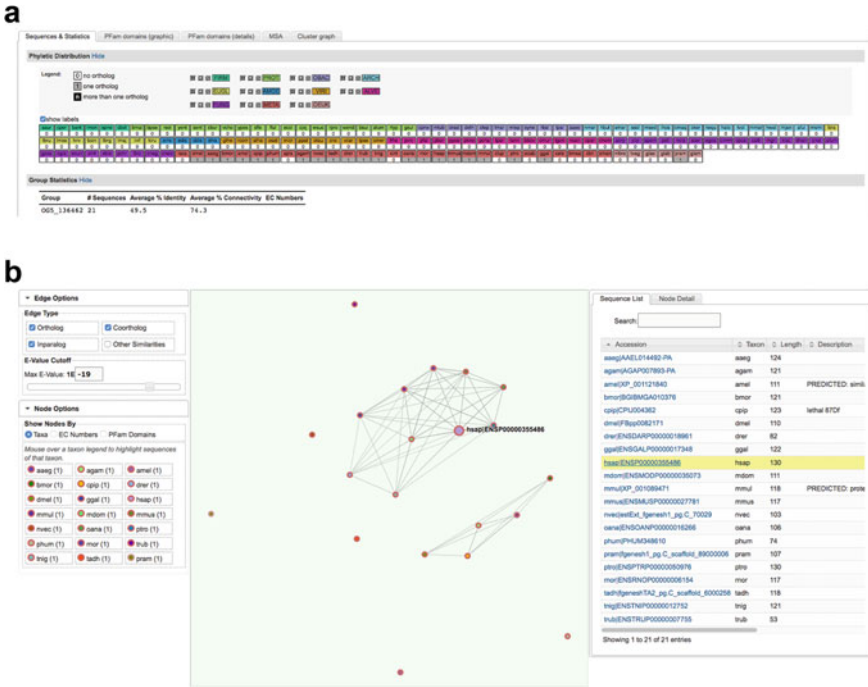
In summary, OMA can be considered a database with a rich visual interface, providing plenty of information and harboring many species. The output is well integrated and visualized. Information on local synteny of a gene of interest should be mentioned, as it is not found in many other orthology databases and can lead to the discovery of gene clusters implied in one mechanism. This is specifically relevant for bacterial proteins. However, OMA was neither able to identify the COX20 orthologs from human and *S. cerevisiae*, nor was it correctly classifying PC with budding yeast PYC1 and PYC2. Nonetheless, these two were correctly identified as close paralogs. Finally, the tailless family was divided in separate, orthologous groups. OMA is available at: <https://omabrowser.org/oma/home/>.

### 3.7 OrthoMCL DB

OrthoMCL DB (Li et al. 2003; Chen et al. 2006; Fischer et al. 2011) is a part of the EuPathDB project and relies on the OrthoMCL clustering algorithm to identify orthologs. Orthologs are identified using WU-BLASTP (Altschul and Gish 1996) and the RBH strategy, using an e-value better than  $1e-5$  as a cut-off for identifying orthologs. Retained orthologs—as well as inparalogs—are linked in a network of orthologs. Edges connecting nodes (orthologs) are weighted using BLAST similarity scores. A graph-based cluster algorithm (Markov Cluster algorithm (MCL)) (Enright et al. 2002) are used to create groups of orthologs. In brief, MCL performs random walks on graphs using Markov matrices to calculate transition probabilities from one node to the other. This graph-based clustering algorithm is less computationally expensive than tree-based methods for clustering orthologs. OrthoMCL DB contains 150 species (36 bacteria, 16 archaea and 112 eukaryotes) and was last updated in July 2015.

The search can be done by OrthoMCL DB IDs, free text search, a phyletic pattern, by function, by groups or by sequence. Searching for the synonym of human COX20, FAM36A, 21 orthologs were found mostly in metazoans. This information is displayed in a simple, colored tabular format on the results page, where abbreviations of species name are associated with a 0 (not found in this species) or a 1 (found in this species) (Fig. 3.9a). The ortholog in *S. cerevisiae* is not found. An interesting feature





**Fig. 3.9** OrthoMCL DB results. **a** Results for a search with human COX20. Most orthologs are found in Metazoans. **b** Graph representation of the COX20 ortholog groups. Individual nodes represent proteins, edges between them represent their orthologous relationship. The graph can be manipulated by choosing different parameters in the control panel on the left-hand side

of OrthoMCL DB is the display of orthologous groups as a network (Fig. 3.9b). This allows interactive visualization of orthologs and how they are related to each other. More or less stringent cut-offs can be chosen to reconstruct the orthology graph. When searching with the text term COX20, 16 orthologs are found, mostly in fungi (eukaryotes). The ortholog in *H. sapiens* is not found for the fungal groups. Information about taxon, identifiers and domain architecture is provided for each gene and species.

Pyruvate carboxylase is classified with other carboxylases. Three proteins are found for *S. cerevisiae* and *H. sapiens*, respectively. These include in human PC, a propionyl-CoA carboxylase and a methylcrotonoyl-CoA carboxylase; in budding yeast, PYC1, PYC2 and an urea amidolyase is included. These different proteins can be subdivided in different orthologous groups. In case of PC, the Enzyme Commission (EC) number can be used to identify the correct enzyme. The EC number of pyruvate carboxylase is 6.4.1.1, meaning it is a part of the ligases (6), forming carbon-carbon bonds (6.4) so ligases that form carbon-carbon bonds (6.4.1); thus, it is a pyruvate carboxylase (6.4.1.1).

The tailless family is split in separate groups, one encompassing tll and NR2E1 and the second one containing Hr51 and NR2E3. We could not find an OrthoMCL group for dsf with any of the valid identifiers, gene names or synonyms.

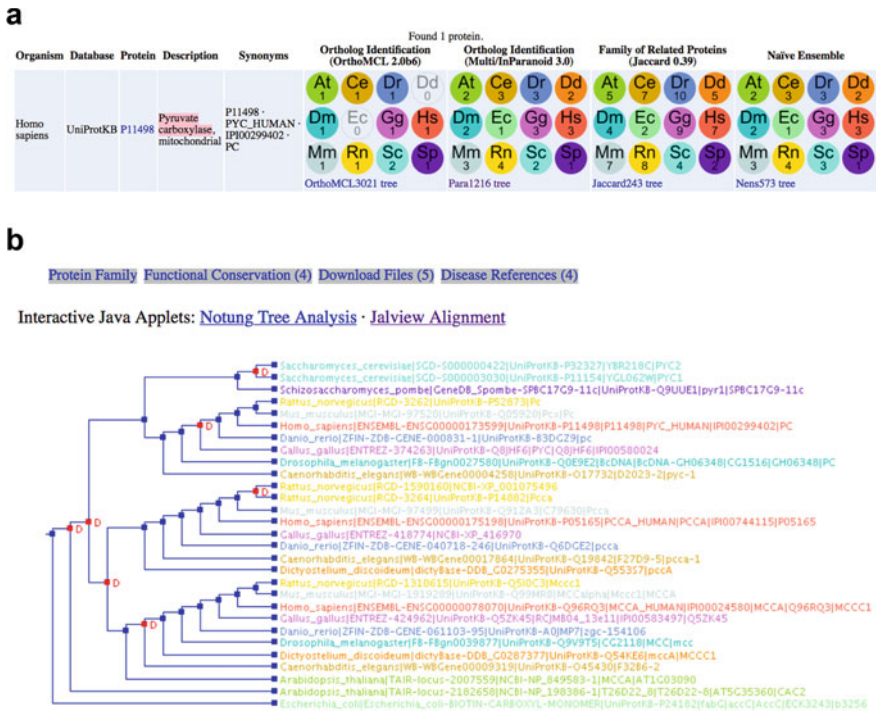
In conclusion, while OrthoMCL provides a fast, graph-based algorithm to cluster orthologs derived from RBHs, the algorithm is not able to identify distant orthologs, nor is it useful when searching for proteins that are part of large superfamilies. Inparalogs are classified in OrthoMCL but, with the implemented clustering algorithm, different orthologs can be clustered in superfamilies, which can be disturbing for a user who is looking for a specific protein. As also observed in other orthology resources, the tailless family is divided in two separate groups. OrthoMCL is available at: <https://orthomcl.org/orthomcl/>.

### 3.8 P-POD

Princeton Protein Orthology Database (P-POD (Heinicke et al. 2007; Livstone et al. 2011)) uses all-against-all BLASTp searches followed by two alternative types of clustering methods to group orthologous proteins: predicted orthologs based on the OrthoMCL clustering algorithm; and larger protein families that are clustered based on a Jaccard index (Jaccard 1912) inferred from shared sequence similarity, putting an orthologous group in its larger evolutionary context. Multiple sequence alignments and phylogenetic trees are created for each group using MAFFT (Katoh and Standley 2013) and PhyML (Guindon et al. 2010), and gene loss and duplication events are resolved using Notung (Chen et al. 2000). Information from species-specific databases are collected for genes, organisms and diseases. It offers information for 12 model organisms including *H. sapiens*, *S. cerevisiae* or *Arabidopsis thaliana*.

Only a text search is possible, using either gene name, IDs or an Online Mendelian Inheritance of Man (OMIM) ID. A search with COX20 returns the yeast protein, which is classified as a sequence orphan. When searching for the synonym of human COX20, FAM36A, P-POD found orthologs in different species, including *Drosophila melanogaster*, but not in *S. cerevisiae*. Looking for pyruvate carboxylase in *H. sapiens* allows to find the two inparalogs in *S. cerevisiae*. According to the method used, different results are found. Using Jaccard clustering, seven proteins are found in *H. sapiens* and four in *S. cerevisiae*, which include other members of this superfamily (Fig. 3.10a). A tree representation can be accessed, showing the evolution of the gene and the possible events of duplication as calculated by the Notung package (Chen et al. 2000) (Fig. 3.10b).

Results for the tailless family show inconsistencies with other resources. There are two distinct P-POD groups for tll (with human NR2E1) and Hr51 (with human NR2E3). The Multi/InParanoid- and naïve Ensemble- Hr51 groups both contain additionally the *Drosophila* Hr83 protein. The Multi/InParanoid group of tll contains in addition the *Drosophila* dsf protein, as well as two human proteins that are not considered part of this family in other resources, namely NR2F1 and NR2F2. The



**Fig. 3.10** P-POD results for pyruvate carboxylase. **a** Using the Jaccard algorithm, four proteins are found for *S. cerevisiae* and seven in *H. sapiens*, while OrthoMCL-based clustering retrieves one ortholog in human and the two inparalogs in budding yeast. **b** Tree representation of the P-POD results. Duplication events are displayed by red squares

naïve Ensemble group of tll contains *Drosophila* proteins tll, dsf and seven-up (svp), as well as the human proteins NR2E1, NR2F1 and NR2F2.

In conclusion, P-POD is less comprehensive than other orthology databases. It only is available for a limited number of model organisms. Due to the stringency of its algorithm, it does not find remote orthologs. The tll family in P-POD includes moreover a number of false-positive proteins. However, as it includes Notung as one step in its analysis pipeline, it is well suited to resolve gene losses and duplications and thus to correctly identify inparalogs. P-POD is available at: <http://ppod.princeton.edu/>.

### 3.9 InParanoid

InParanoid (O'Brien et al. 2005; Sonnhammer and Östlund 2015) uses reciprocal BLASTp searches to identify orthologs via the RBH method. In Version 8, there are 273 proteomes in the database (246 eukaryotes, 20 bacteria and seven archaea),

extracted from Ensembl and UniProt. Inparalogs are separated in the output, and outparalogs are excluded. The user can set the score for excluding inparalogs when invoking an InParanoid search.

InParanoid offers several different search options, including text search, identifier search, or a sequence-based search. When searching for COX20 in human, all pairwise groups of orthologs are returned, which makes navigation of results difficult (Fig. 3.11a). The budding yeast COX20 ortholog is not found. Likewise, using the human COX20 sequence for the search, *S. cerevisiae* COX20 is not found. Searching for FAM36A even in a text search returns orthologs of human COX20, however not the human protein itself, which indicates that alternative identifiers are not supported. When searching for COX20 in *S. cerevisiae*, only its orthologs in other fungi are found.

InParanoid is designed for identifying inparalogs, so it is not surprising to find members of the pyruvate carboxylase correctly (Fig. 3.11b). Using either PC or PYC is non-practical as text search, as the database cannot disambiguate the name of this gene. Searching for the gene name PC returns mostly clusters of pyruvate carboxylase, yet includes also polycomb protein c (Pc) from *Drosophila melanogaster*.

Like HomoloGene, InParanoid groups *Drosophila* dsf with human NR2E1. Tll is in a separate InParanoid cluster, not containing a human ortholog. Hr51 is clustered with NR2E3.

In conclusion, InParanoid offers a moderate range of organisms and is limited to well-conserved orthologs that can be found by BLASTp. Inparalogs are resolved correctly, and the user can choose a score to include or exclude inparalogs. The tailless family is split in three clusters, whereby dsf is considered orthologous to NR2E1, not tll. The output of InParanoid is pragmatic and simple, however non-practical, as each cluster the query is found in, is shown separately and no family cluster is created for proteins belonging to the same orthologous group. InParanoid is available at: <http://inparanoid.sbc.su.se/cgi-bin/index.cgi>.

### 3.10 KEGG Orthology Database

The KEGG orthology database (Kanehisa et al. 2014, 2016a, 2017) is a part of the Kyoto Encyclopedia of Genes and Genomes (KEGG). It contains at least 4000 genomes. Orthology data are collected using KOALA (KEGG Orthology And Links Annotation (Kanehisa et al. 2010)). This tool evaluates similarity scores, best-hit relationships, domains and taxonomy to assign genes and proteins to a group of orthologs. The data are also verified manually using experimental evidence and literature. The orthologs are classified in groups by a specific KOALA number (K number). The orthology groups are fully integrated with the rest of the KEGG resources, for example, by linking to KEGG Genes. KOALA groups are linked to BRITE hierarchies and KEGG pathway maps. The KOALA group can also be displayed as a simple hierarchy, following the NCBI taxonomic classification.

**a**

Inparalog and Orthologs cluster for *Homo sapiens* and *Drosophila melanogaster*


Cluster 5199					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q5R115	<i>Homo sapiens</i>	1	100%	Cytochrome c oxidase protein 20 homolog	COX20_HUMAN (UniProt)
Q9VVG00	<i>Drosophila melanogaster</i>	1	100%	Lethal (3) 870f	Q9VVG00_DROME (UniProt)

**b**

Inparalog and Orthologs cluster for *Homo sapiens* and *Saccharomyces cerevisiae*

Cluster 16					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
P11498	<i>Homo sapiens</i>	1	100%	Pyruvate carboxylase, mitochondrial	PYC_HUMAN (UniProt)
P11154	<i>Saccharomyces cerevisiae</i>	1	100%	Pyruvate carboxylase 1	PYCL_YEAST (UniProt)
P32327	<i>Saccharomyces cerevisiae</i>	0.883		Pyruvate carboxylase 2	PYCC_YEAST (UniProt)

**c**



**ORTHOLOGY: K18184**

<b>Entry</b>	K18184	KO
<b>Name</b>	COX20	
<b>Definition</b>	cytochrome c oxidase assembly protein subunit 20	
<b>Pathway</b>	ko04714	Thermogenesis
<b>Disease</b>	H01368	Cytochrome c oxidase (COX) deficiency
<b>Brite</b>	KEGG Orthology (KO) [BR:ko00001] 09150 Organismal Systems 09159 Environmental adaptation 04714 Thermogenesis K18184 COX20; cytochrome c oxidase assembly protein subunit 2 09180 Brite Hierarchies 09182 Protein families: genetic information processing 03029 Mitochondrial biogenesis K18184 COX20; cytochrome c oxidase assembly protein subunit 2 Mitochondrial biogenesis [BR:ko03029] Mitochondrial quality control factors Mitochondrial respiratory chain complex assembly factors Complex-IV assembly factors K18184 COX20; cytochrome c oxidase assembly protein subunit 2 <a href="#">BRITE hierarchy</a>	
<b>Genes</b>	HSA: 116228(COX20) PTR: 736354(COX20) PPS: 100992118(COX20) GGO: 101143171 PON: 100442587 NLE: 100606669 MCC: 100425230 704298(COX20) MCF: 102122147 102125892 CSAB: 103218044 103230894(COX20) RRO: 104654195 » show all <a href="#">Taxonomy</a>   <a href="#">KOALA</a>   <a href="#">UniProt</a>	
<b>Reference</b>	PMID:10671482	
<b>Authors</b>	Hell K, Tzagoloff A, Neupert W, Stuart RA	
<b>Title</b>	Identification of Cox20p, a novel protein involved in the maturation and assembly of cytochrome oxidase subunit 2.	
<b>Journal</b>	J Biol Chem 275:4571-8 (2000) DOI:10.1074/jbc.275.7.4571	
<b>Sequence</b>	[scc:YDR231C]	

**All links**

- Ontology (2)
- KEGG BRITE (2)
- Pathway (2)
- KEGG PATHWAY (2)
- Disease (1)
- KEGG DISEASE (1)
- Gene (625)
- KEGG GENES (358)
- KEGG MOGENES (5)
- RefGene (255)
- OC (7)
- Protein sequence (198)
- UniProt (193)
- SWISS-PROT (5)
- Literature (1)
- PubMed (1)
- All databases (829)
- Download RDF

**Fig. 3.11** Simplistic results of the InParanoid and KEGG orthology database. **a** InParanoid results for COX20. As *S. cerevisiae* was not found, we show an example with *D. melanogaster*. **b** InParanoid results for pyruvate carboxylase. The inparalogs are grouped with their ortholog in *H. sapiens*. **c** KEGG results for COX20. Next to the basic information on the protein, the main information provided comes from the Brite annotations. The list of orthologs is shown in the main box, as are potential articles associated with an ortholog group. Links to all other resources from KEGG are given in the ‘All links’ box

The remote ortholog of *S. cerevisiae* is found for human COX20. These proteins belong to the large group K18184 with over 300 members. The orthology information is displayed in KEGG style, with Brite functional annotation of the orthologous group. COX20 for instance belongs to the group 09150 associated with Organismal Systems and the group 04714 associated with thermogenesis. The genes that are part of group K18184 are listed right below the BRITE hierarchy and links to the KEGG organisms within the NCBI taxonomy, the KOALA list of genes, as well as the UniProt list of genes. The last part of the box contains information on the literature and the sequence entry itself. Next to the main box, all links within KEGG are shown (Fig. 3.11c).

The name ambiguity of pyruvate carboxylase (PC) again gave a long list of results, whereas looking for PYC gave only five results. Thus, it is best to search for pyruvate carboxylase or to use an identifier accepted by KEGG. PC belongs to the group K01958, and the two *S. cerevisiae* inparalogs are correctly identified. The EC (IUBMB) has assigned the EC number 6.4.1.1.

KEGG separates the tailless family in three KO groups. Tll is clustered with human NR2E1, Hr51 with human NR2E3. Finally, dsf does not have a human ortholog.

In conclusion, KEGG orthology is a highly interlinked database, which can take advantage of all information available by KEGG including information on genes, pathways, ontologies, disease and literature. The database is not very visual, except for the available pathway maps, which makes browsing the results somewhat difficult. Its annotation strategy however correctly identified the remote ortholog COX20 from *S. cerevisiae* and is also able to resolve inparalog relationships. The tailless family is split into three groups, with the orthology assignments following the predominant consensus of other orthology resources. KEGG orthology database is available at: <https://www.genome.jp/kegg/ko.html>.

### 3.11 EggNOG

Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (EggNOG) was developed by the Computational Biology team at the EMBL in Heidelberg (Jensen et al. 2008; Huerta-Cepas et al. 2019). We describe here the latest available pipeline, which consists of many steps. Reciprocal hits are derived from all-against-all Smith–Waterman alignments provided by the SIMAP project (Arnold et al. 2014). In the next step triangular clustering—searching for reciprocal best hits using sets of three species—is performed to identify orthologous groups (OG). In this pipeline, inparalogs are identified and treated as one sequence to ensure that they finally belong to the same cluster. Each OG is annotated using a functional annotation pipeline that consolidates the annotations of the identified species within an OG. Similar to OrthoDB, different taxonomic levels are used to compute OGs independently. This ensures a more accurate functional annotation. These nested OGs are tested and corrected for consistencies. A phylogenetic tree is calculated for each OG using the Python-based ETE pipeline (Huerta-Cepas et al. 2016). In brief, multiple sequence

alignments are created using Clustal Omega (Sievers et al. 2011) and soft trimmed to remove columns with low coverage; ModelFinder (Kalyaanamoorthy et al. 2017) is used to test the model, and a maximum likelihood tree is generated using IQ-TREE (Nguyen et al. 2015). The current version, 5.0.0, contains 7562 organisms: 4445 bacteria, 168 archaea, 447 eukaryotes and 2502 viruses.

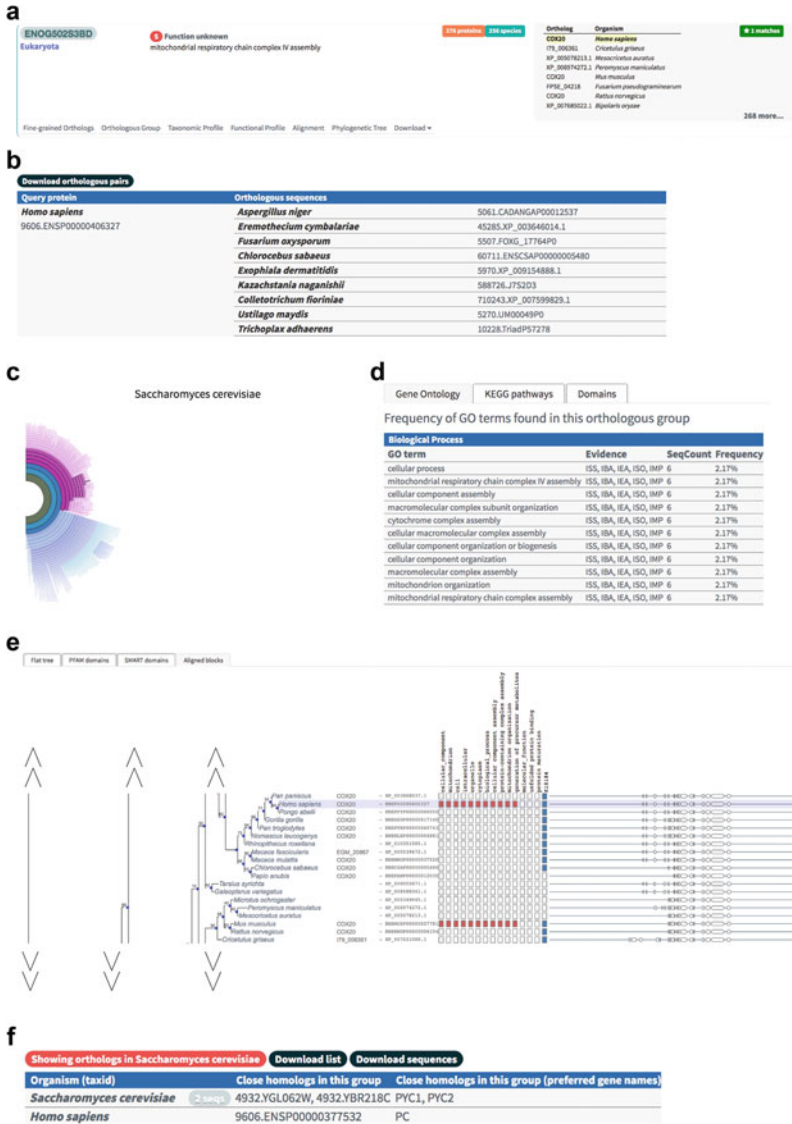
A search is divided in multiple steps. The user has to first enter a search term (e.g., COX20), then select a species and indicate, in which taxonomic range the orthologous group should be searched in. When searching for COX20 and adding *H. sapiens* in EggNOG 5, the OG group ENOG502S3BD is found, containing 276 proteins found in 256 species, which indicates that inparalogs are included in OGs in EggNOG (Fig. 3.12a). The budding yeast COX20 ortholog is found in EggNOG v5, whereas it was not found in the previous version of the tool (4.5.1). Orthologs are listed with their identifiers from different databases, such as NCBI and Ensembl (Fig. 3.12b). The results include different tabs showing a taxonomic profile (Fig. 3.12c), GO terms, KEGG pathway, and conserved domains (Fig. 3.12d), a multiple alignment of all the orthologs (trimmed and untrimmed), as well as a phylogenetic tree that is decorated with functional annotation and conserved domains (Fig. 3.12e). All information for the protein family, including a Hidden Markov Model (HMM) consisting of all family members, is downloadable from the site. When searching for pyruvate carboxylase, 524 proteins are found in 367 species, the two inparalogs of *S. cerevisiae* were correctly found and are clustered together.

EggNOG provides an overview of the orthologous group tailless belongs to, which is called the steroid hormone mediated signaling pathway family. It also classifies the fine-grained, pairwise orthologs in this family. Tll, dsf and Hr51 all belong to the same EggNOG orthologous group. *Drosophila* has in total 14 members in this group, human 38. Tll itself is the pairwise ortholog of NR2E1, Hr51 of NR2E3. Finally, dsf has no human ortholog.

In conclusion, the use of EggNOG is quite easy, and it can find remote orthologs and inparalogs. It moreover gives the complete overview of the orthologous group of tailless, which contains a number of paralogs in the different species, together with providing a detailed view on pairwise orthologs within this family. It applies a hierarchical procedure to cluster orthologs and offers a rich set of information and visualization of OGs. EggNOG is actively maintained and is available at: <http://egg-nog5.embl.de/#/app/home>.

## 3.12 PANTHER

PANTHER (Thomas et al. 2003) is part of the Gene Ontology Phylogenetic Annotation Project, led by the Gene Ontology consortium (Gaudet et al. 2011). This acronym stands for Protein ANalysis THrough Evolutionary Relationships. It contains 142 genomes, 35 bacteria, 8 archaea, 99 eukaryotes. Currently, version 15.0 is online, updated in February 2020. PANTHER's main goal is to provide high-confident functional annotations by classifying proteins according to their evolutionary history.



**Fig. 3.12** EggNOG information provided for ortholog groups, here for COX20. **a** EggNOG group for COX20 at the Eukaryota level with a brief functional description, as well as the number of proteins and organisms found (top boxes). **b** An exhaustive list of orthologs is given, linking to different source databases to retrieve the sequence (restricted for illustration purposes). **c** the taxonomic profile of COX20. The profile is interactive and the user can browse at different levels of the taxonomic hierarchy. Here shown for *S. cerevisiae* (levels passed are shown in this case in dark pink). **d** GO terms associated with the COX20 group, providing information on the evidence of a GO term. **e** Phylogenetic tree that is decorated with aligned segments of the ortholog group. **f** EggNOG results for PC, with the two identified inparalogs, PYC1 and PYC2, in *S. cerevisiae*



Next to providing information on protein families and pathways, PANTHER also offers its own, reduced (i.e., slim) ontology for functional categorization. Since version 7, orthologs are annotated within PANTHER. PANTHER infers orthologs from family trees, based on pairs of genes who have diverged by a speciation event. Families are first separated based on the PANTHER HMM library. Multiple sequence alignments from these families are constructed using MAFFT (Katoh and Standley 2013), which are then used for tree reconstruction using Giga (Thomas 2010). Not only one-to-one, but also one-to-many orthologs are inferred, reporting for instance inparalogs. In case of one-to-many relationships, PANTHER also reports the least diverged orthologs, which are believed to still have the same function.

When searching for human COX20 in ‘genes and orthologs,’ a list of all orthologs is returned, with links to their entries in PANTHER (Fig. 3.13a) giving ample information, such as IDs from other databases and alternate IDs (Fig. 3.13b), PANTHER families and subfamilies the gene belongs to (Fig. 3.13c), PANTHER GO and GO slim annotations, as well as all orthologs of the gene (Fig. 3.13a). An interactive phylogenetic tree (full and reduced) can be displayed, which can be annotated with the full or trimmed multiple sequence alignment, including information on evolutionary events, such as deletions, insertions or mutations (Fig. 3.13 d). Other types of information displayed include identifiers, family and subfamily associations, or PANTHER IDs of functional annotations.

Next to searching in ‘genes and orthologs,’ PANTHER families can also be searched. The term COX20 returns three families, whereby two correspond to the cytochrome c oxidase assembly protein COX20. Both are part of the same family and represent two different subfamilies, SF1 and SF2. The SF2 subfamily contains the human COX20 and three orthologs from primates, whereas the SF1 subfamily contains the budding yeast protein and all other orthologs from a wide range of different species, including *Mus musculus*, *Danio rerio* or *Candida albicans*. The PANTHER family for COX20 (PTHR31586) contains a third subfamily, SF4, which is not named and which contains only *Macaca mulatta* and *Pan troglodytes*. An HMM of each family can be downloaded. Looking for pyruvate carboxylase (or PYC) in human, only one subfamily is found (PTHR43778). In fact, it is the only PANTHER family for this protein and consists of 73 genes. The two PYC genes of *S. cerevisiae* as well as human PC are in this PANTHER family.

PANTHER distinguishes the gene family and subfamilies for the nuclear hormone receptors tll, dsf and Hr51. All three are part of the nuclear hormone receptor (NHR) family, which includes 517 proteins from the supported 33 species and which are presented in the accompanying family tree at the PANTHER website. Tll, dsf and Hr51 are also part of their own subfamily, whereby tll groups with human NR2E1, Hr51 with NR2E3 and dsf is considered arthropod-specific. Gene counts of the NHR family, however, differ between EggNOG and PANTHER: While EggNOG includes 14 *Drosophila* and 38 human members for this family, PANTHER only considers eight *Drosophila* and 12 human proteins as part of the family.

Taken together, PANTHER is easy to handle, finds remote orthologs and gives information about inparalogs. Like EggNOG, PANTHER shows both the entire family of nuclear hormone receptors for the tll family and the subfamilies with the



**Fig. 3.13** PANTHER information on the COX20 family. **a** List of orthologs of COX20 with links to their entries in PANTHER. **b** PANTHER gene information about COX20, linking via identifiers to different databases. **c** Classification of the gene in PANTHER and the PANTHER slim ontology. **d** Interactive phylogenetic tree, annotated with the multiple sequence alignment used to calculate the tree, including information on evolutionary events, such as deletions, insertions or mutations

consistent grouping of tll with NR2E1 and Hr51 with NR2E3. The division of organisms from the same orthologous group in subfamilies can be confusing. PANTHER is available at: <http://www.pantherdb.org/>.

All orthology resources that we discussed here are summarized in Table 3.2, indicating their performance (with respect to their potential to detect remote orthologs, or classify inparalogs), the possibility of programmatic access, as well as data availability.

### **3.13 Do-It-Yourself: Availability of Search Algorithms and Orthology Data from Orthology Resources**

Most of the databases discussed provide their search algorithm and/or pre-calculated orthology data for download and local usage. This is specifically useful if users want to annotate their own genome, or place a large number of proteins in orthologous groups. On the other hand, pre-calculated orthologs can be useful for data mining or large-scale phylogenetic analyses.

#### ***3.13.1 Programmatic Access and Data Download***

Programmatic access, such as Application Programming Interfaces (APIs), is available for most resources. These interfaces allow users to download data from the database's website within their own pipelines and on a large scale. All databases except for HCOP, P-POD and InParanoid allow programmatic access to their data. Data from Homologene, HCOP or OMA additionally can be accessed directly via R.

All databases except for KEGG allow also download of their data, including multiple sequence alignments, HMM-libraries and phylogenetic trees, if available, which is useful for large-scale phylogenetic data mining. In order to avoid compatibility problems of formats, orthoXML (Schmitt et al. 2011) is used to store and compare orthology data from a wide range of databases, which is for instance offered by HCOP, InParanoid, OMA, OrthoMCL or PANTHER.

#### ***3.13.2 Availability of Code***

Several orthology inference tools are also downloadable as a stand-alone version. They can be used to identify orthologs and orthology groups of newly sequenced genomes assisting in the proper functional annotation of genes and proteins, phylogenetic profiling, or species tree reconstruction.

**Table 3.2** Summary of the tools studied

Tools	Method	Number of species	(B)acteria/(A)rchaea/(E)ukaryota	Remote orthologs	Inparalogs	Programmatic access	Data download	References
OrthoDB	RBH + clustering	13,772	B/A/E	✓	✓	✓	✓	Eygenia et al. (2019)
HomoloGene	RBH + Tree	21	E	✗	✓	✓	✓	N.C.B.I Resource Coordinators (2013)
TreeFam	Tree	109	E	✗	✓	✓	✓	Ruan et al. (2008)
HCOP	Combination of databases	19	E (human-centric)	✓	✓	✗	✓	Eyre et al. (2007)
OMA	Smith and Waterman + clustering	2288	B/A/E	✗	✓	✓	✓	Altenhoff et al. (2018) Altenhoff et al. (2015)
OrthoMCL	Markov Cluster algorithm + graph flow theory	150	B/A/E	✗	✓	✓	✓	Li et al. (2003) Chen et al. (2006)
P-POD	Combination of databases + Tree	12	E	✗	✓	✗	✓	Heinicke et al. (2007)

(continued)

Table 3.2 (continued)

Tools	Method	Number of species	(B)acteria/(A)rchaea/(E)ukaryota	Remote orthologs	Inparalogs	Programmatic access	Data download	References
InParanoid	RBH + clustering	273	B/A/E	✗	✓	✗	✓	O'Brien Kevin et al. (2005), Sonnhammer and Östlund (2015)
KEGG Orthology	KOALA + Manual curation	>4000	B/A/E	✓	✓	✓	✗	Kanehisa et al. (2016a, b) Kanehisa et al. (2014)
EggNOG	Smith and Waterman + clustering + tree + HMM	7562	B/A/E	✓	✓	✓	✓	Huerta-Cepas et al. (2019)
PANTHER	HMM + tree	142	B/A / E	✓	✓	✓	✓	Thomas et al. (2003)

Given information relies on the version available in January 2020

The pipeline used by OrthoDB can be downloaded for stand-alone usage. It represents the full pipeline used by the OrthoDB resource. After solving some dependencies for multithreading and boosting, the ORTHOPIPE and BRHCLUS software packages are easy to install and test locally. The stand-alone pipeline is available from the OrthoDB website (<https://www.orthodb.org/?page=software>).

An OMA stand-alone version is downloadable and usable as a command line tool (Altenhoff et al. 2019). Installation and usage instructions are well-written and easy to follow; moreover, parallelization instructions are provided to run larger OMA-jobs. The software package is very easy to install. OMA lists four major areas of application for its stand-alone version: species tree reconstruction, genome annotation, dynamics of genome evolution (making use of the HOG clusters) and finally phylogenetic profiling, looking for gene absences or duplications. Output is provided in OrthoXML format, as well as FASTA and tabular files. OMA is available for download from the OMA website (<https://omabrowser.org/standalone/#downloads>), as well as from GitHub (<https://github.com/DessimozLab/OmaStandalone/blob/master/OMA.drw>).

OrthoMCL is available as a downloadable stand-alone version. Next to a local BLAST, it depends on a local installation of a relational database (MySQL or Oracle), for storing the orthologous pairs and orthology groups, as well as the MCL software for graph-based clustering. Installation instructions are available. The pipeline consists of individual perl scripts that need to be executed one after the other. Estimated run-time is given, which indicates the necessity of a larger compute cluster to run OrthoMCL locally. The software has only been tested for RedHat 5.8. The stand-alone version is available from the OrthoMCL website (<https://orthomcl.org/common/downloads/software/v2.0/>). MySQL, the MCL cluster software, as well as BLAST need to be downloaded and installed prior to installation of the pipeline. A new SQLite-dependent pipeline, which also contains a wrapper-script, is available from GitHub: (<https://github.com/stajichlab/OrthoMCL>).

InParanoid is available as a stand-alone version for calculating pairwise orthologs, as well as orthologs among 3 organisms. It depends on NCBI-BLAST. Currently, InParanoid only supports the old version of BLAST and does not support the blast+ package, which is now standard. The user needs to either have a compatible version of BLAST available, or manipulate the InParanoid perl program to work with the newer blast+.

KEGG offers the BlastKOALA, GhostKOALA and KofamKOALA programs to automatically assign genome sequences to K numbers (KO assignment), however only as an online tool (Kanehisa et al. 2016b). The user can upload sequences for mapping to KEGG Orthology groups. Three different search algorithms are used: standard BLAST (BlastKOALA, <https://www.kegg.jp/blastkoala/>), or GHOSTX (GhostKOALA, <https://www.kegg.jp/ghostkoala/>), which is a fast homology search algorithm relying on query and database suffix arrays for seed matching. In the newest version of the KEGG search tools, an HMM profile-based search algorithm, HMMER3, (KofamKOALA (Aramaki et al. 2020) (<https://www.genome.jp/tools/kofamkoala/>)) is used. KofamKOALA searches against a pre-computed database of HMMs derived from KO families.

EggNOG offers the EggNOG mapper online (<http://eggnog-mapper.embl.de/>) and for download (<https://github.com/eggnogdb/eggnog-mapper>) to functionally annotate entire proteomes based on orthology. The stand-alone version can be easily cloned from GitHub and is easy to install. Sufficient documentation is provided, which is equally easy to follow. EggNOG mapper is also available online for annotating novel proteomes.

Finally, PANTHER provides the set of their tools for download at <http://pantherdb.org/downloads/index.jsp>, as well as on GitHub (<https://github.com/pantherdb>). It includes the PANTHER HMM scoring tool, which allows to compare a set of sequences, e.g., from a newly sequenced genome, against the entire PANTHER HMM library. The PantherScore tool depends on HMMER3, which needs to be installed independently. Among the tools provided is also the Java-based PAINT tree viewer program, as well as db-PAINT (from the GitHub repository) that allows functional annotation based on phylogenetic analysis. Installation instructions are given and easy to follow. Some essential information is missing; for example, it is unclear how the taxonomic ID file should be structured and how to retrieve it from NCBI.

### 3.14 Discussion

There exist many databases and tools that help identify orthologous genes or groups. We have presented the most commonly used and known available resources. We however do not claim that our list is exhaustive. We have tested the databases whether they can find remote orthologs and how they deal with paralogous genes, more precisely with inparalogs as well as with families with complex evolutionary histories. We found that all databases were able to handle inparalogs correctly. However, only few of them contained information on the remote ortholog we were searching for. Among those were OrthoDB, HCOP, KEGG, EggNOG and PANTHER. We would like to note at this point that while those resources were able to find COX20, we do not have any further data to support the claim that they contain all possible remote orthologs. The nuclear hormone receptor family tailless was particularly difficult to place for many resources. This is not surprising, as this family shows a large expansion in some taxonomic phyla, such as nematodes. While most of the databases were able to still correctly limit this group to the core members of *D. melanogaster* and *H. sapiens*, many of them contained putative false-positive members from *C. elegans*. Particularly, TreeFam and P-POD fail to correctly classify this family. TreeFam, for example, contains proteins from *C. elegans* that are considered orthologous to other nuclear hormone receptor families in other resources: *C. elegans* unc-55 is for instance orthologous to NR2F1 from *H. sapiens*. It is furthermore noteworthy that two orthology databases, HomoloGene and InParanoid, define dsf as the ortholog of human NR2E1, instead of tll, even though it is not the best reciprocal hit, but only the second-best hit. This can be explained by the higher number of identical amino acids

in local alignments between dsf and NR2E1 and the 239 amino acid long insertion in the middle of the dsf protein, which renders the RBH results ambiguous.

Different databases typically use different identifiers. Most accept gene names as a search item. Yet, ambiguous gene names such as the official gene symbol for pyruvate carboxylase, PC, find too many hits in the databases. Navigating search results can therefore be problematic. A text-based search for the full name, or using an identifier accepted by the database resolved this problem.

Orthology databases differ in the availability of additional annotation provided for orthologous groups. In this respect, databases embedded in genome resources have an advantage, as the entire information collected on genes is easily available. These include for instance HomoloGene, PANTHER or KEGG.

Some databases do not contain up-to-date information. This means in most cases that no new species were integrated in the database. While we do not see this as a reason for not making use of a resource, it indicates that curation might have been neglected for these databases.

Among the databases tested, we found that OrthoDB was one of the databases with the largest number of available organisms. It also had the most complete set of orthologs, including remotely conserved orthologs. It has linked all entries to several databases, provides domain annotation, as well as functional annotation of orthologous groups. Moreover, the NCBI gene resource meanwhile relies for orthology information next to HomoloGene also on OrthoDB. We see the hierarchical treatment of orthologous groups employed by OrthoDB, but also by other resources like EggNOG as an advantage. Functions are more likely retained in closely related species, and thus, making use of a more fine-grained, taxonomical clustering will result in more accurate functional annotation transfer.

The availability of code as well as Web-based services is required for annotation of newly sequenced genomes. Many resources allow users to either download their code locally or provide Web-based services for whole-genome annotations. While we have not tested each available tool locally, we found that they are easy enough to install and usage instructions are easy to follow. To run such computationally heavy tools locally can however exceed dramatically the resources available to a user, both in compute time, as well as in storage space. When considering orthology-based functional assignment, this should be kept in mind. Using Web-based services like EggNOG, KEGG or PANTHER provides a viable solution to this problem.

Finally, we want to stress the importance of adding newly sequenced genomes to orthology search pipelines. First of all, novel model organisms arise rapidly, for instance to address questions in evolutionary developmental biology. Providing a good functional annotation of sequences by knowledge transfer from orthologs will help advance scientific discovery in non-standard model organisms. Second, adding new species will lead to a better coverage of search space to find orthologs. This will ultimately also help in discovering remote orthologies and in gaining a better understanding of the evolution of genes and pathways.

**Acknowledgements** We would like to thank the support staff from the NCBI for help with data retrieval and for providing information; Maria-Mandela Prunster, Amélie Vernale and Friedhelm



Pfeiffer for critical reading of the manuscript. This work was supported by the CNRS and by an MENRT thesis grant from the French Ministry of Research awarded to Paul de Boissier.

## References

- Altenhoff AM, Glover NM, Train C-M et al (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* 46:D477–D485. <https://doi.org/10.1093/nar/gkx1019>
- Altenhoff AM, Levy J, Zarowiecki M et al (2019) OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res* 29:1152–1163. <https://doi.org/10.1101/gr.243212.118>
- Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* 8:e1002514. <https://doi.org/10.1371/journal.pcbi.1002514>
- Altenhoff AM, Škunca N, Glover N et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res* 43:D240–D249. <https://doi.org/10.1093/nar/gku1158>
- Altschul SF, Gish W (1996) Local alignment statistics. *Meth Enzymol* 266:460–480. [https://doi.org/10.1016/s0076-6879\(96\)66029-7](https://doi.org/10.1016/s0076-6879(96)66029-7)
- Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402. <https://doi.org/10.1093/nar/25.17.3389>
- Aramaki T, Blanc-Mathieu R, Endo H et al (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36:2251–2252. <https://doi.org/10.1093/bioinformatics/btz859>
- Arnold R, Goldenberg F, Mewes H-W, Rattei T (2014) SIMAP—the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage. *Nucleic Acids Res* 42:D279–D284. <https://doi.org/10.1093/nar/gkt970>
- Bondy JA, Murty USR (1976) *Graph theory with applications*. North Holland
- Brown D, Sjölander K (2006) Functional classification using phylogenomic inference. *PLoS Comput Biol* 2:e77. <https://doi.org/10.1371/journal.pcbi.0020077>
- Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34:D363–D368. <https://doi.org/10.1093/nar/gkj123>
- Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol* 7:429–447. <https://doi.org/10.1089/106652700750050871>
- Cunningham F, Achuthan P, Akanni W et al (2019) Ensembl 2019. *Nucleic Acids Res* 47:D745–D751. <https://doi.org/10.1093/nar/gky1113>
- El-Gebali S, Mistry J, Bateman A et al (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427–D432. <https://doi.org/10.1093/nar/gky995>
- Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584. <https://doi.org/10.1093/nar/30.7.1575>
- Eyre TA, Wright MW, Lush MJ, Bruford EA (2007) HCOP: a searchable database of human orthology predictions. *Brief Bioinformatics* 8:2–5. <https://doi.org/10.1093/bib/bbl030>
- Federhen S (2012) The NCBI Taxonomy database. *Nucleic Acids Res* 40:D136–D143. <https://doi.org/10.1093/nar/gkr1178>
- Fischer S, Brunk BP, Chen F et al (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics Chapter 6:Unit 6.12*. 1–19. <https://doi.org/10.1002/0471250953.bi0612s35>

- Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinformatics* 12:449–462. <https://doi.org/10.1093/bib/bbr042>
- Guindon S, Dufayard J-F, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321. <https://doi.org/10.1093/sysbio/syq010>
- Heinicke S, Livstone MS, Lu C et al (2007) The Princeton protein orthology database (P-POD): a comparative genomics analysis tool for biologists. *PLoS ONE* 2:e766. <https://doi.org/10.1371/journal.pone.0000766>
- Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 33:1635–1638. <https://doi.org/10.1093/molbev/msw046>
- Huerta-Cepas J, Szklarczyk D, Heller D et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47:D309–D314. <https://doi.org/10.1093/nar/gky1085>
- Jaccard P (1912) The Distribution of THE flora in the Alpine zone. I. *New Phytol* 11:37–50. <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>
- Jensen LJ, Julien P, Kuhn M et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 36:D250–D254. <https://doi.org/10.1093/nar/gkm796>
- Kalyaanamoorthy S, Minh BQ, Wong TKF et al (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589. <https://doi.org/10.1038/nmeth.4285>
- Kanehisa M, Furumichi M, Tanabe M et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45:D353–D361. <https://doi.org/10.1093/nar/gkw1092>
- Kanehisa M, Goto S, Furumichi M et al (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38:D355–D360. <https://doi.org/10.1093/nar/gkp896>
- Kanehisa M, Goto S, Sato Y et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42:D199–205. <https://doi.org/10.1093/nar/gkt1076>
- Kanehisa M, Sato Y, Kawashima M et al (2016a) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–D462. <https://doi.org/10.1093/nar/gkv1070>
- Kanehisa M, Sato Y, Morishima K (2016b) BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* 428:726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>
- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780. <https://doi.org/10.1093/molbev/mst010>
- Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–338. <https://doi.org/10.1146/annurev.genet.39.073003.114725>
- Kriventseva EV, Kuznetsov D, Tegenfeldt F et al (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 47:D807–D811. <https://doi.org/10.1093/nar/gky1053>
- Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* 36:D271–D275. <https://doi.org/10.1093/nar/gkm845>
- Li H, Coghlan A, Ruan J et al (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* 34:D572–D580. <https://doi.org/10.1093/nar/gkj118>
- Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189. <https://doi.org/10.1101/gr.1224503>
- Livstone MS, Oughtred R, Heinicke S et al (2011) Inferring protein function from homology using the Princeton Protein Orthology Database (P-POD). *Curr Protoc Bioinformatics* Chapter 6:Unit 6.11. <https://doi.org/10.1002/0471250953.bi0611s33>
- Lu S, Wang J, Chitsaz F et al (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* 48:D265–D268. <https://doi.org/10.1093/nar/gkz991>

- NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44:D7–19. <https://doi.org/10.1093/nar/gkv1290>
- NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 46:D8–D13. <https://doi.org/10.1093/nar/gkx1095>
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>
- O'Brien KP, Remm M, Sonnhammer ELL (2005) InParanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 33:D476–D480. <https://doi.org/10.1093/nar/gki107>
- Pronk JT, Yde Steensma H, Van Dijken JP (1996) Pyruvate metabolism in *Saccharomyces cerevisiae*. *Yeast* 12:1607–1633. [https://doi.org/10.1002/\(sici\)1097-0061\(199612\)12:16%3c1607::aid-yea70%3e3.0.co;2-4](https://doi.org/10.1002/(sici)1097-0061(199612)12:16%3c1607::aid-yea70%3e3.0.co;2-4)
- Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277. [https://doi.org/10.1016/s0168-9525\(00\)02024-2](https://doi.org/10.1016/s0168-9525(00)02024-2)
- Ruan J, Li H, Chen Z et al (2008) TreeFam: 2008 Update. *Nucleic Acids Res* 36:D735–D740. <https://doi.org/10.1093/nar/gkm1005>
- Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinformatics* 12:485–488. <https://doi.org/10.1093/bib/bbr025>
- Schreiber F, Patricio M, Muffato M et al (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res* 42:D922–D925. <https://doi.org/10.1093/nar/gkt1055>
- Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539. <https://doi.org/10.1038/msb.2011.75>
- Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43:D234–D239. <https://doi.org/10.1093/nar/gku1203>
- Steinegger M, Meier M, Mirdita M et al (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20:473–515. <https://doi.org/10.1186/s12859-019-3019-7>
- Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026–1028. <https://doi.org/10.1038/nbt.3988>
- Szklarczyk R, Wanschers BFJ, Nijtmans LG et al (2013) A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia. *Hum Mol Genet* 22:656–667. <https://doi.org/10.1093/hmg/dds473>
- Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:1–14. <https://doi.org/10.1186/1471-2105-4-41>
- Thomas PD (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics* 11:312–319. <https://doi.org/10.1186/1471-2105-11-312>
- Thomas PD, Campbell MJ, Kejariwal A et al (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13:2129–2141. <https://doi.org/10.1101/gr.772403>
- Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34:1692–1699. <https://doi.org/10.1093/nar/gkl091>
- Walter F (1989) R. F. Doolittle, Of URFS and ORFS—a primer on how to analyze derived amino acid sequences. VII + 103 S., 24 Abb., 14 Tab. Mill Valley 1986. University Science Books. ISBN: 0-935702-54-7. *J Basic Microbiol* 29:246–246. <https://doi.org/10.1002/jobm.3620290411>
- Wolf YI, Koonin EV (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol* 4:1286–1294. <https://doi.org/10.1093/gbe/evs100>
- Zahn-Zabal M, Dessimoz C, Glover NM (2020) Identifying orthologs with OMA: a primer. *F1000Res* 9:27. <https://doi.org/10.12688/f1000research.21508.1>

# Chapter 4

## Protein Recoding Through RNA Editing: Detection, Function, Evolution



Eli Eisenberg

**Abstract** RNA editing is an endogenous post-transcriptional process that alters the RNA sequence, changing its information content from that encoded in the DNA. Throughout the animal kingdom, the most common type of RNA editing is A-to-I editing, catalyzed by double-stranded RNA-specific adenosine deaminase (ADAR) enzymes. ADARs mostly target non-coding RNA sequences. However, some protein-coding regions are modified, resulting in non-synonymous substitutions and novel protein products. These editing sites, also known as “recoding” sites, contribute to the complexity and diversification of the proteome. Computational transcriptomic studies have identified thousands of recoding sites in multiple species, many of which are conserved within lineages. However, the functional impact of recoding, in most cases, is yet to be revealed. In this chapter we discuss the utility of recoding for diversity and adaptation throughout evolution.

### 4.1 Introduction

RNA editing is a post-transcriptional modification that alters the information content of the RNA sequence itself (Bass 2002; Nishikura 2016; Eisenberg and Levanon 2018). Across metazoa, the most prevalent type of RNA editing is adenosine to inosine (A-to-I) editing, mediated by members of the well-conserved ADAR (adenosine deaminase acting on RNA) enzyme family. Two catalytically active enzymes of this family are encoded in the mammalian genome: ADAR1 (also known as ADAR) and ADAR2 (also known as ADARB1). ADAR1 is strongly expressed in all tissues (Lonsdale et al. 2013). ADAR2 expression is lower than that of ADAR1. It is expressed most highly in the artery, cerebellum, esophagus and lung tissues, although observed to some extent in most other tissues as well (Lonsdale et al. 2013).

---

E. Eisenberg (✉)

Raymond and Beverly Sackler School of Physics and Astronomy and Sagol School of Neuroscience, Tel Aviv University, 69978 Tel Aviv, Israel

e-mail: [elieis@post.tau.ac.il](mailto:elieis@post.tau.ac.il)

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_4](https://doi.org/10.1007/978-3-030-57246-4_4)

79

ADARs were first identified as enzymes that unwind double-stranded RNA (dsRNA) structures (Rebagliati and Melton 1987; Bass and Weintraub 1988). It is now widely believed that this dsRNA unwinding function is the ancestral function of the widely expressed ADAR1 protein, accounting for the lethal phenotype of ADAR1 deletion in mice (Hartner et al. 2009; Mannion et al. 2014; Liddicoat et al. 2015; Pestal et al. 2015; George et al. 2016). Long double-stranded RNAs (dsRNAs) are identified by sensor proteins such as MDA5, and trigger production of type I interferons as part of recruiting the innate immunity system against viral RNA (Schneider et al. 2014; Wu and Chen 2014). However, large numbers of endogenous dsRNAs are likely to appear in normal eukaryotic cells as well (Reich and Bass 2019), mainly due to the abundance of mobile elements in the genome—transcripts harboring nearby inverted copies of the same repeat fold to create an endogenous dsRNA structure (Porath et al. 2017b). These structure may erroneously trigger the cytosolic immune response, resulting in a severe outcome to the host cell (Hartner et al. 2009; Mannion et al. 2014; Liddicoat et al. 2015; Pestal et al. 2015; George et al. 2016). A-to-I editing, mostly carried out by the constitutive ADAR1p110 variant, introduces mismatches to the endogenous dsRNAs while still in the nucleus (Patterson and Samuel 1995; Roth et al. 2019), so that the edited endogenous transcripts are no longer recognized by dsRNA sensors in the cytoplasm, possibly through destabilization of the RNA structure. Preventing the endogenous dsRNAs from false alarming the immune system is the essential function of ADAR1.

In parallel with editing and unwinding the potentially dangerous long and nearly perfect dsRNAs, ADARs also edit much shorter and weaker structures. Many such structures are bound to appear in the transcriptome, due to the abundance of repetitive elements. In fact, all multicellular metazoans screened so far (Porath et al. 2017a, b) exhibit extensive editing, the extent of which strongly depends on the repertoire of repetitive elements in their genome (Neeman et al. 2006; Porath et al. 2017b). Likely, most of this extensive editing is not crucial for preventing an innate immune response (Barak et al. 2020).

The vast majority of this editing activity occurs in non-coding regions, such as the primate-specific *Alu* repetitive elements (Levanon et al. 2004), and is catalyzed mostly by ADAR1 (Roth et al. 2019). In some cases, noncoding editing events may have acquired a function. For example, the cellular fate of an mRNA and/or its translation probability can be affected by editing of miRNA binding sites in its 3' UTR (Pinto et al. 2017) or by editing of the cognate miRNAs themselves (Kawahara et al. 2007; Alon et al. 2012; Vesely et al. 2012; Pinto et al. 2017; Wang et al. 2017). Yet, as of now it seems that most of these sites are functionally irrelevant.

The situation is quite different with respect to the coding sequence. Due to the structural similarity, inosines mimic guanines in many cellular processes (Basilio et al. 1962). Translation of inosine-containing codons is mostly similar to that of the equivalent guanosine-containing ones (except for IAC codons, where 25% of the translated proteins interpret the inosine as adenosine) (Licht et al. 2019). Thus, editing of protein-coding sequences may lead to non-synonymous substitutions and novel protein variants, possibly affecting protein functionality. In addition to point-like protein modifications, editing may create splice sites, resulting in the introduction

of novel exons (Rueter et al. 1999; Lev-Maor et al. 2007), and editing of a stop codon (e.g., UAG (stop)  $\rightarrow$  UIG (tryptophan)) may lead to stoploss and C-terminal extension of the protein. Thus, unlike non-coding editing, the functional potential of editing events modifying the resulting protein (“recoding” sites) is quite clear. In mammals, recoding sites are mainly targeted by ADAR2, and it is thus believed that the main function of the mammalian ADAR2 enzyme is to edit specific non-synonymous sites within protein-coding sequences (Tan et al. 2017).

Although other types of RNA editing also lead to recoding, to the best of our knowledge A-to-I editing is the only type that gives rise to recoding in nuclear mRNA across multiple tissues and conserved across lineages. The rest of this chapter focuses on A-to-I recoding.

## 4.2 Observing Recoding in RNA-Seq Data

Most current RNA sequencing schemes start with reverse transcription of the RNA into cDNA. Like ribosomes, reverse-transcriptases treat the inosines as guanosines. Consequently, inosines in the mRNA appear as guanosines in the cDNA, and the editing events show up in the RNA-seq data as A-to-G DNA-RNA mismatches.

Discovery of the first mammalian recoding sites throughout the first decade of A-to-I RNA editing research were serendipitous. The introduction of computational approaches has enabled systematic large-scale editing detection. The basic idea behind these approaches is quite simple. As editing shows up as an A-to-G DNA-RNA mismatch, one only needs to scan through large-scale sequencing databases and look for this mismatches, filtering out technical and biological noise (e.g., sequencing errors, incorrect alignment, genomic polymorphisms, somatic mutations) (Eisenberg et al. 2010; Schrider et al. 2011; Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012; Piskol et al. 2013). Since 2003, a number of groups have developed computational approaches that apply various filters to the multitude A-to-G mismatches observed in a given sample, or a set of samples, in order to identify the relatively few originating from an editing event (Levanon and Eisenberg 2006; Eisenberg 2012; Ramaswami and Li 2016; Diroma et al. 2017; PMID: 32211029 Claudio Lo Giudice et al, “Quantifying RNA Editing in Deep Transcriptome Datasets”). Advances in sequencing technologies have increased the availability of high coverage multi-sample datasets, resulting in millions of editing sites identified in human and other species (Bazak et al. 2014a; Ramaswami and Li 2014; Picardi et al. 2017a).

These systematic searches revealed that recoding is but an exception of the editing repertoire. Virtually all sites found in the abovementioned computational screens reside out of the coding region and have no direct effect on the protein. Furthermore, non-coding editing events are easier to find, as they are often clustered and concentrated in the well-identified repetitive elements. As a result, on top of the low numbers of recoding sites detected, the false-positive rate is very high in the coding region, especially for mammalian transcriptomes where the scope of recoding is rather low compared with *Drosophila* or cephalopods (see below).

Accordingly, standard widely used all-purpose detection schemes are not suitable for detection of recoding events. While they do show an impressive transcriptome-wide performance, the results in coding regions are rather poor (as reflected by the low fraction of A-to-G mismatches among all mismatches found). The reliability of thousands of putative human recoding sites that have been reported by the large-scale systematic searches for editing sites is thus questionable. Reliable identification of recoding sites is yet an unmet challenge.

One effective approach is available for conserved recoding sites. The technical and biological errors mentioned above are not expected to reoccur in multiple species at the exact same location, and therefore conserved A-to-G mismatches that are observed at the same position in two (not-too-close) species are expected to be enriched in evolutionarily conserved recoding sites (Hoopengardner et al. 2003; Levanon et al. 2005; Pinto et al. 2014). Note, however, that in highly conserved exons one may observe the same alignment artifact in several species, leading to a false discovery of a “conserved recoding event.” Dedicated methods for detection of recoding events in a single-species data are being developed currently. Hopefully, a conservative alignment that minimizes alignment errors supplemented by utilization of multiple samples to filter out genomic polymorphisms may be the key to reliable and comprehensive mapping of recoding sites.

## 4.3 Utility of Recoding

### 4.3.1 *Diversifying the Proteome*

Recent decades have revealed the important role played by post-transcriptional and post-translational mechanisms in generating the proteomic complexity of higher organisms. These epigenetic mechanisms allow for diversification of the proteome in a temporally regulated, tissue-specific, condition-dependent way, leading to functional heterogeneity across tissues, developmental stages, brain regions or even among individual cells within the same tissue.

Recoding by A-to-I RNA editing is an example for such a mechanism, facilitating proteome diversification. It has the capacity to create a range of proteins from a single genomically encoded gene, providing the organism with a new means for acclimation and adaptation. Unlike genomic mutations, editing could modify a fraction of the transcript copies, and its levels may be fine-tuned to produce the edited and unedited versions of the protein concurrently, even within the same single cell, at a relative concentration that depends on the tissue, condition and environment. Indeed, several studies have demonstrated how recoding levels at specific sites do change as a function of the organism’s condition. For example, editing in a variety of transcripts was shown to modulate along the circadian cycle of transcripts in mouse

liver (Terajima et al. 2016), and changes in RNA editing have been associated with sleep (Robinson et al. 2016). Importantly, many studies have demonstrated altered editing of individual recoding targets in various disease states [for a recent review, see (Gallo et al. 2017)].

Recoding facilitates a much wider range of possibilities for adjusting the transcriptome than genomic mutations do. Unlike genomic mutations, the edits are transient, well-suited to respond immediately to external cues and drive acclimation to changes in internal or environmental conditions, without compromising the genomic information. A nice demonstration of this idea is the peak in ADAR levels and editing levels during spawning in corals, leading to over a thousand recoding events at the time of gamete release that are not observed in adult corals (Porath et al. 2017a). This extensive increase in protein diversity may improve gamete's adaptability without manipulating the underlying genome (Eisenberg and Levanon 2018). Another intriguing example is provided by recoding of a potassium channel in octopus, whose level correlates with the external temperature. It is not yet clear, however, whether this effect is due to rapid acclimation or long-term adaptation (Garrett and Rosenthal 2012a, b).

While recoding probably occurs in virtually all metazoa, the repertoire of recoding sites varies considerably across lineages. Only a few dozen recoding sites are known to be conserved across mammals (Pinto et al. 2014). Similarly, dozens of sites were found in zebra fish (Sie and Maas 2009; Pozo and Hoopengardner 2012; Li et al. 2014a; Shamay-Ramot et al. 2015), ants (Li et al. 2014b), as well as 164 sites in bees (Porath et al. 2019). The situation is somewhat different in *Drosophila*, where nearly a thousand recoding sites were shown to be conserved across the lineage (Yu et al. 2016; Duan et al. 2017; Zhang et al. 2017). The most notable exception is the cephalopod's lineage, utilizing recoding at a level that far surpasses all other species studied so far (Alon et al. 2015; Liscovitch-Brauer et al. 2017), with tens of thousands of recoding events found in each of the four coleoid cephalopod species studied.

### ***4.3.2 Limitations on Functional Utilization of Recoding***

Given the above-described potential of recoding to be functionally utilized, and the fact that the editing mechanism is encoded in the metazoan genome, the relatively limited scope of recoding is surprising. One may have expected that in the course of organisms' evolution, recoding sites will appear and fixate in the transcriptome as a response to external pressures. However, with the exception of cephalopods, recoding seems to be utilized to a rather limited extent across the animal kingdom. Even in *Drosophila* and cephalopods, the contribution of the conserved recoding sites to adaptation is not clear (Yablonovitch et al. 2017a, b). Why would that be the case? Several possible explanations have been proposed.



One possibility is that regulation of RNA editing is not sufficiently complex to allow for individual control of each of the hundreds or thousands of functional recoding sites. As far as is currently known, the editing efficiency is mostly determined by two factors: local sequence and structural motifs encoded in the RNA sequence, and the expression level of the ADAR proteins and their regulators. The surrounding sequence is, by and large, hard-wired in the genome, and is therefore independent on the tissue, cell-type, environmental condition or developmental stage. Indeed, editing levels at specific mammalian sites are largely consistent across tissue-matched samples from different individuals (Greenberger et al. 2010). Thus, this factor does not contribute to regulation, and one would expect the variations in editing level at a given site to be mostly governed by the level of the ADAR proteins and their regulators. Alterations in ADAR levels might allow intricate tissue-dependent or condition-dependent regulation (Picardi et al. 2015), but all editing sites would be equally affected. This sets a major limitation on the flexibility of regulation, and may result in an effective upper bound to the number of independently regulated functional recoding sites.

It should be noted, though, that the full repertoire of ADAR regulators is still unknown. Possibly, there are multiple trans-regulators of RNA editing that allow for a more complex editing pattern (several candidates have been recently suggested Fritz et al. 2009; Marcucci et al. 2011; Garncarz et al. 2013; Behm et al. 2017; Oakes et al. 2017; Tan et al. 2017; Chung et al. 2018; Roth et al. 2019). Note, however, that the enzyme specificity of these regulators is mostly unknown. Possibly they affect mostly ADAR1. Another interesting layer of editing regulation is provided by auto-editing of ADAR2 (Rueter et al. 1999), resulting in the appearance of a novel 3' splice acceptor site, which in turn leads to an addition of 47 nucleotides. The affected transcript is frame-shifted, predicted to lose the dsRNA-binding domain as well as the catalytic domain. Interestingly, ADAR-auto-regulation is also observed in *Drosophila* and bumblebee, but there it leads to non-synonymous changes rather than a frameshift (Palladino et al. 2000; Savva et al. 2012; Porath et al. 2019). However, as far as we currently know these ADAR regulators mostly affect editing globally, and probably do not allow for site-specific control of editing levels. More intricate, yet unidentified, layers of regulation may exist, providing differential control over the editing levels at different sites. On the other hand, if indeed editing regulation, by and large, does not provide site-specific resolution, this sets a major limitation on the use of recoding for adaptation and acclimation. These limitations become more and more pressing with an increasing number of functional recoding sites, as adjustment of the global regulators of recoding should take into account the effect on an increasing number of targets.

Another possible explanation for the rare usage of recoding in many species is related to the evolutionary cost of maintaining a fixed functional recoding site. It has been suggested (Liscovitch-Brauer et al. 2017) that conservation of an active recoding site imposes a severe constraint on the genomic region that encodes the dsRNA structure recognized by ADAR proteins. Mutations that affect the stability of this secondary structure might modify the level of editing or abolish editing altogether (Reenan 2005; Rieder et al. 2013). If the site is indeed positively selected, such

mutations will undergo purifying selection so that the delicate balance between the edited and unedited versions of the protein is maintained. The higher the number of such positively selected sites is, the stronger is this constraint on the global genomic evolution. In cephalopods, it is estimated that 3–15% of the inter-species mutations and 10–26% of the intra-species polymorphisms were purified due to constraints associated with maintenance of editing (Liscovitch-Brauer et al. 2017). Conversely, creation of a new editing site requires a structure to evolve, imposing evolutionary constraints on the surrounding sequence. This trade-off between the transcriptome plasticity provided by RNA editing and the genomic variation required to drive adaptation and evolution might explain why extensive recoding was disfavored in most metazoan lineages (Liscovitch-Brauer et al. 2017).

### ***4.3.3 Recoding as a Global Response to External Conditions***

However, even if recoding cannot be efficiently regulated at a single target resolution, global regulation of recoding may be still useful for adaptation if a change in external conditions, such as temperature or acidity, affects all sites, or many of them, in a similar way. Recoding may then be utilized to counteract this change, or response to it, in all recoding sites. For example, editing has been shown to be involved in temperature response in both *Drosophila* and cephalopods (Garrett and Rosenthal 2012b; Rieder et al. 2015; Buchumenski et al. 2017). Presumably, a decrease in the external temperature perturbs the energy-entropy balance controlling protein-folding and might be mitigated by a global increase in editing that tends to replace multiple amino acids by smaller, less stabilizing, ones (Garrett and Rosenthal 2012a). Under this scenario, global coordinated upregulation of editing in multiple targets could be functional as a response mechanism to lowered temperatures. Interestingly, this response of editing to temperature, one of the most important environmental variables, can be easily achieved without any need in intricate regulatory networks. Editing depends on folding the RNA molecule into dsRNA structures. The stable folded structure is governed by a balance between binding energies and structural entropy, and is therefore affected directly by the external temperature. It is therefore easy to imagine RNA structures that are fine-tuned to allow editing only below a certain cut-off temperature.

Having the above scenario in mind, one is tempted to offer an attractive explanation to the striking difference between mammals on one side, and *Drosophila* and cephalopods on the other. The latter species have been shown to utilize recoding to respond to acute temperature changes, while the homeothermic mammals have no incentive to utilize extensive recoding. This is further supported by a recent study that examined RNA editing in squirrel, a heterothermic mammal, and suggested a dynamic response of the A-to-I editing profile to the low body temperature during

hibernation (Riemondy et al. 2018). One should note, however, that the above-mentioned initial analyses of ants, bees, and fish, seem to suggest that limited-scope recoding is not limited to homeothermal animals. Future studies of more diverse species are needed to reveal the extent to which cold-blooded organisms utilize extensive editing to respond to temperature.

#### 4.3.4 *Functional Studies of Specific Sites*

The previous sections leave us with a number of open questions: Is RNA editing utilized for proteome diversifications? If so, which of the editing events is adaptive? Is conserved recoding generally adaptive? Does editing contribute to a dynamic proteomic response to external pressures? Detailed functional analyses of multiple recoding sites are required in order to fully settle these questions. However, experimental studies of the effect of recoding are often challenging and time-extensive, as the phenotype of editing may be subtle, if not elusive. Accordingly, mechanistic understanding of the effect of recoding in these sites on the biochemical activity of the protein, not to mention functional analysis of the consequences to the cell and the organism, typically lags behind identification of new recoding sites. So far, only some of the strongly edited and conserved mammalian sites have been characterized in detail.

The most studied recoding site is the Q/R site in GluR-B, the first discovered case of recoding in mammals, which results in voltage-independent gating with decreased calcium permeability (Sommer et al. 1991; Higuchi et al. 1993; Seeburg and Hartner 2003). Editing of this site is nearly complete in normal brain tissues (Sommer et al. 1991). Its under-editing is associated with human diseases such as amyotrophic lateral sclerosis (ALS) and malignant gliomas (Maas et al. 2001; Kawahara et al. 2004; Kwak and Kawahara 2005) and the absence of recoding at this site results in an early death in mice (Higuchi et al. 2000). This is the only mammalian recoding site associated with such a severe phenotype. The Q/R site is one of the most conserved recoding sites in mammals, observed in amphibians and some species of fish, and is likely to have been evolved no later than the appearance of cartilaginous fish (Kung et al. 2001).

The second target identified, the serotonin 2C receptor (Burns et al. 1997) (5-HT<sub>2c</sub>R) is one member of a family of serotonin receptors expressed in the central nervous system, edited in five different sites affecting three amino acids. These sites are not fully edited, nor fully correlated, and thus editing could potentially lead to 24 different protein isoforms with varying effect on the response to serotonin and a cascade of downstream pathways (Burns et al. 1997; Marion et al. 2004). Transcripts encoding for at least 20 of the different protein variants were observed in human brain tissues (Wang et al. 2000; Wahlstedt et al. 2009; Khermesh et al. 2016; Zaidan et al. 2018). However, the unedited isoform (Isoleucine–Asparagine–Isoleucine; INI) alone accounts for roughly half of the transcripts (Khermesh et al. 2016).

Functional studies of the effect of recoding have been published for a small number of other physiologically important mammalian genes (Sommer et al. 1991; Egebjerg and Heinemann 1993; Lomeli et al. 1994; Burns et al. 1997; Sailer et al. 1999; Bhalla et al. 2004; Yeo et al. 2010; Daniel et al. 2011; Chen et al. 2013; Miyake et al. 2016; Jain et al. 2018), and electrophysiological studies have analyzed the effects of recoding on a few ion channels in cephalopods (Patton et al. 1997; Rosenthal and Bezanilla 2002; Colina et al. 2010; Liscovitch-Brauer et al. 2017), but the implications of recoding remain largely unknown for the vast majority of reported sites.

Over one thousand recoding sites reported in humans, but only a few dozen of them were shown to be conserved across mammals (Pinto et al. 2014). Thus, the vast majority of human recoding sites seem to be restricted to human or the primate lineage. These non-conserved recoding sites do not show signs of selection (Xu and Zhang 2014)—that is, they are less abundant and more weakly edited compared with editing at synonymous sites, and they are under-represented in essential genes, highly expressed genes, and genes that are under purifying selection. However, it is not clear yet whether these results represent the actual behavior of mammalian recoding sites or merely reflect the rather large false-positive rate in current databases.

Furthermore, even for the conserved sites the functional importance of editing is not obvious. A recent study has demonstrated that, with the exception of the essential recoding Q/R site within *GRIA2* transcripts (Higuchi et al. 2000), complete abolishment of recoding is well tolerated (Chalk et al. 2019). Mice lacking ADAR2 suffer from progressive seizures and die within three weeks of birth, but this severe phenotype is completely rescued by altering their genome to encode an arginine at the *GRIA2* Q/R recoding site (Higuchi et al. 2000; Chalk et al. 2019). The rescued mice develop normally and live a normal lifespan even if ADAR1-editing is further shut down (Chalk et al. 2019). This unexpected result does not exclude the possibility that recoding of conserved mammalian targets (other than the Q/R *GRIA2* site) does have functionally important, even if subtle (Horsch et al. 2011) (or apparent only under specific conditions), effects. However, it raises the possibility that many of these sites may be dispensable.

Finally, the vast majority of the mammalian recoding sites reported so far are edited to a very low level. Often, only a few percent or less of the transcripts carry the edited version. Certainly, low-level editing is less likely to have a functional impact. Indeed, the editing levels at the conserved recoding sites, expected to be adaptive, are much higher than that of the non-conserved sites, or the synonymous editing sites with the coding sequence (Pinto et al. 2014). Assuming the low-level sites are not functional, why are they being edited? This may be just a biological noise, as ADAR enzymes may bind weakly to some randomly structured RNAs and edit them to a minimal extent. In parallel, many weakly edited sites are due to “satellite” editing. The RNA structures required for editing of functionally important recoding sites often include dozens, or even hundreds, of adenosine nucleotides. Some of these may get edited just because they happen to be incorporated in the dsRNA structure. In both cases, these events may survive selection as long as the effect of editing is not too deleterious (e.g., editing is weak enough so that the slight decrease in the

unedited protein isoform is tolerable and the edited form itself is not harmful) (Xu and Zhang 2014). Satellite sites may even be conserved across distant species, as a result of conservation of the structure required for editing of the functional site in their vicinity.

However, it is also possible that sites appearing to be weakly edited when averaged over a tissue, exhibit much higher editing levels in specific subpopulations of cells (Gal-Mark et al. 2017), or even at a single-cell level (Picardi et al. 2017b). In fact, an interesting recent report suggests that at the single-cell level, editing is often binary in nature—either all copies of the transcript are being edited, or none are (Picardi et al. 2017b). If this is indeed the case, then even a low-level of editing could have a major impact on some cells within the tissue.

## 4.4 Evolutionary Aspects of Recoding

### 4.4.1 *The Evolutionary History of Recoding*

The ancestral ADAR enzyme appears to have originated via the incorporation of a double-stranded RNA binding domain into the coding sequence of ADAT1, a member of the ADATs family (adenosine deaminases acting on tRNA) found in all eukaryotes (Gerber et al. 1998) that are incapable of editing mRNAs. Extensive editing has been observed in cnidaria (corals) (Porath et al. 2017a), and ADAR enzymes were identified in multiple Ctenophora and Porifera species (although not in the placozoan *Trichoplax adhaerens*) suggesting that the origin and expansion of the ADAR gene family preceded the last common ancestor to all contemporary animals (Grice and Degan 2015). It is now widely believed that the ancestral function of ADAR1, shared by all present-day metazoans, is to protect against false activation of the innate immune system. Recoding is probably a secondary use of the editing machinery. Following the introduction of ADARs to the metazoan cell, weak recoding sites have presumably appeared as a side-effect to the ancestral ADAR1 activity, and the beneficial ones were then maintained and further evolved.

It should be noted that while the RNA edits themselves are transient and are not transmitted to the next generation of cells, editability is inherited through the RNA structural and sequence motifs encoded in the parent genomic sequence. As editing relies on the target RNA adopting a specific dsRNA secondary structure, and possibly adjacent editing-enhancing dsRNA structures (Lomeli et al. 1994; Rieder et al. 2013; Daniel et al. 2014; Sapiro et al. 2015), the genomic sequence surrounding a sites may transmit the editing pattern to the next generation of cells, and genomic mutations in this sequence may further fine-tune editing efficiency. Recoding is therefore a mechanism for heritable proteome diversification and has the potential to lead to adaptation in response to external pressures (Gommans et al. 2009).

A novel recoding site may appear in the course of evolution following an accumulation of random point mutations that slowly modify the structure of the corresponding RNA molecule to form of the minimal dsRNA structure required for ADAR recruitment. This process may be accelerated by the activity of mobile elements, in two different ways. First, mobile elements newly integrated to the genome may be exonized and incorporated into protein-coding sequences (Sorek et al. 2004). These repetitive elements are susceptible to editing, as they can readily pair with a similar reversely oriented element in a nearby intron to create a long and stable dsRNA duplex (Bazak et al. 2014b). For example, the hundreds of *Alu* elements that have been exonized into coding regions of the human transcriptome (Dagan et al. 2004) are enriched in primate-specific recoding sites (over a thousand such sites are tabulated in current databases). A notable example is the *NARF* gene, harboring a pair of extensively edited inverted *Alu* repeats in one of its introns. In primates, editing of *NARF* pre-mRNA creates a novel splicing site and recodes a stop-codon, resulting in a novel primate-specific alternatively spliced exon, which itself contains additional recoding sites (Lev-Maor et al. 2007).

Second, mobile elements may accelerate the emergence of novel recoding events by creating an intronic RNA duplex as a result of mobile element activity in a nearby intron. Long and stable intronic dsRNAs are known to induce or enhance site-selective editing at recoding sites in a neighboring exon, up to several hundred nucleotides away (Daniel et al. 2012, 2017; Ramaswami et al. 2015). Notably, many of the most efficiently edited (>50% editing) recoding sites conserved across mammals are located in proximity to a nearby editing-inducing elements (Daniel et al. 2017) that may serve as ADAR recruitment elements. Accordingly, a pair of inverted mobile elements newly introduced near a coding exon could form a dsRNA structure that would enhance editing of a neighboring preexisting recoding site, or even initiate recoding at a site that was not edited prior to insertion of the repetitive element (Daniel et al. 2014).

Interestingly, the genetic code prevents the appearance of a premature stop codon due to an adenosine into guanosine substitution. Thus, random non-specific A-to-I editing events cannot produce truncated protein products, usually dysfunctional and often harmful, and their potential deleterious effect is limited. This observation may partially explain how extensive A-to-I editing is tolerated (as compared to C-to-U editing, for example). Most nonspecific recoding is expected to be evolutionarily neutral or slightly deleterious and should be slowly depleted from the transcriptome, while the few beneficial sites are fixed. If this model is correct, one may expect to see in present-day transcriptomes many newly acquired recoding sites that are organism-specific (or lineage-specific) and mostly evolutionarily neutral or possibly mildly deleterious, in addition to a set of more deeply conserved, functionally beneficial, fixed sites.

Indeed, virtually all recoding sites identified in mammals, *Drosophila*, cephalopods, and other species studied so far are lineage-specific, and most of them are not conserved even across closely related species. Thousands of human recoding sites have been reported, only a few dozens of which were found in mouse, and only a handful are known to be edited in non-mammalian vertebrates. For example,

editing of the Q/R site in GluR-B is observed in birds, amphibians and some species of fish, assumed to have been acquired following the Agnatha–Gnathostome separation (Kung et al. 2001), and recoding of FLNA and CYFIP2 is conserved in birds (Levanon et al. 2005). So far, only a single target (the Shaker potassium channel) is known to be shared by vertebrates, *Drosophila* and cephalopods (Porath et al. 2019). Thus, while the available information about the conservation of recoding across species is still partial, it seems consistent with the view that recoding sites were not part of the ancestral set of ADAR targets, but rather were exapted into the genomes of the different lineages subsequent to their divergence, possibly following a lineage-specific large-scale genome invasion of mobile elements. Screening of more lineages is then expected to reveal independent sets of recoding sites, of widely varying size.

#### 4.4.2 *Interplay Between Recoding and Genomic Mutations*

Interestingly, many recoding sites are fixed genomically as guanosines in closely related species (Tian et al. 2008; Pinto et al. 2014). In some cases, the ancestral genomic allele is G, and then editing partially counteracts the effect of a G-to-A genomic mutation. For example, it is argued that the Q/R site in GluR-B has emerged following the divergence of jawed vertebrates. The ancestral allele, as appears in jawless fish (but also in many teleost fish, including zebra fish and fugu) codes for arginine (Kung et al. 2001). Similarly, frog and puffer fish genomic versions of subunit  $\alpha 3$  of the GABA<sub>A</sub> receptor encode for methionine at a position orthologous to the mammalian-conserved I/M recoding site (Ohlson et al. 2007). In these cases, one may argue that the genomic-A allele is disadvantageous, and it is only due to editing that the G-to-A mutation can be tolerated and fixated. If this is the case, recoding should have evolved rather quickly (on evolutionary scales, obviously) following the genomic G-to-A conversion, which means that the mutation should have occurred within a pre-existing dsRNA structure. It is yet to be determined whether in such cases having the “editing switch,” i.e., the possibility to express both the edited and non-edited variants of the protein, is beneficial compared with having only the edited version hard-wired G in the genome.

On the other hand, there are several examples for sites where the ancestral genomic state was an editable adenosine, and then in some species a guanosine was hardwired into the genome. For example, one of the recoding sites in subunit  $\alpha 6$  of the nicotinic acetylcholine receptor is recoded in the silkworm and the honeybee, but the tobacco budworm harbors a genomically encoded G (Tian et al. 2008). Phylogenetic analysis reveals that the ancestral state at this site is an adenosine, which has gained recoding in some species, and then was converted to a guanosine in the tobacco budworm. In such cases, it is tempting to think of editing as an evolutionary intermediate, enabling “probing” of the G allele without changing the genome. Only when the organism is well-adjusted to the G allele, can the genomic A-to-G mutation be accepted (Tian

et al. 2008). However, currently available data is limited to anecdotal examples and can be equally explained by the simple observation that sites where the G allele is tolerable are more likely to acquire both recoding and a genomic A-to-G mutation.

### 4.4.3 *Is Recoding Generally Adaptive?*

What fraction of recoding activity is adaptive? Analysis of thousands of human putative recoding sites suggests that these sites are mostly non-adaptive and slightly deleterious (Xu and Zhang 2014). Only a few dozen human coding sites are conserved across mammalian species (Pinto et al. 2014) and expected to be functional. The situations seems very different in other lineages: close to a thousand recoding sites are conserved across the *Drosophila* lineage (Yu et al. 2016; Duan et al. 2017; Zhang et al. 2017), as well as more than 10,000 recoding sites conserved across cephalopod species (Liscovitch-Brauer et al. 2017). These sites show signs of positive selection and are enriched for non-synonymous substitutions (recoding sites) over synonymous substitutions, an indicator of positive selective pressure.

Even in mammals, the question of recoding adaptiveness is not fully settled. First, it is not yet clear to what extent these analyses are affected by the high false-discovery rates in the reported sites. An improved analysis of the adaptive nature of recoding in mammals requires a more accurate detection scheme, as well as a more detailed analysis of conservation in closer species, e.g., within the primate lineage. Second, as explained above, many weak editing sites are expected to arise due to nonspecific ADAR activity, so adaptiveness should be analyzed based on the editing levels. In fact, although these weak sites are numerous, their overall contribution to the recoding activity (measured by the number of deamination reactions) is not large compared to the conserved sites that are strongly expressed and strongly edited. In most human tissues, recoding of *FLNA* and *IGFBP7*, whose recoding is both conserved across mammals and has a proven functional impact (Jain et al. 2018; Morgantini et al. 2019), accounts for the majority of ADAR's recoding deamination reactions. Thus, while it may very well be the case that most recoding sites are nonadaptive, most recoding activity may be adaptive. Third, some weak sites are "satellite" events that belong to a cluster of sites including a stronger, possibly conserved and functional site. The latter sites may be nonadaptive standing alone, but editing of the whole cluster may still be beneficial.

On the other hand, the adaptive role of conserved recoding activity was recently challenged from a different angle (Jiang and Zhang 2019). It was suggested that editing as a diversifying mechanism is actually never adaptive, and the only cases in which editing is conserved and maintained by evolution are those where only the G allele is actually beneficial. According to this "harm-permitting model," recoding is fixated in the genome only when required to correct for a deleterious G-to-A genomic mutation ("restorative editing," which may be the case for the Q/R *GRIA2* site, see above), or at least to compensate for the lack of a beneficial A-to-G mutation. One may argue that such cases are not truly adaptive, as having a fixed G allele would be



advantageous over the flexible editable adenosine. Restorative non-adaptive editing may account for the over-representations of recoding sites (high  $N/S$ , nonsynonymous to synonymous ratio) observed in conserved mammalian sites, as well as *Drosophila* and cephalopod sites, even if there is no adaptive advantage to having an editable A at these sites as compared to the ancestral genomically encoded G. This “harm-permitting” model is supported by analysis of cephalopods’ recoding sites exhibiting enrichment of recoding in restorative ancestral-G sites, consistently with prior studied (Tian et al. 2008; Zhang et al. 2014; An et al. 2019). While restorative editing certainly takes place, its extent is still unclear. It is not known yet whether it may account for the multitude of deeply conserved sites. Careful analysis of the evolutionary history of recoding sites in multiple lineages and experimental analysis of known conserved sites are required in order to settle this fundamental and important question.

## 4.5 Conclusion

Recoding is a post-transcriptional mechanism, capable of diversifying the proteome and contributing to its complexity. Despite much progress in the past three decades, a number of key basic questions are still open. Computational biologists are still struggling to provide comprehensive and accurate sets of recoding sites, even in human. On the experimental side, the biochemical and functional impact of recoding is largely unknown for the majority of the strongly edited and well-conserved sites. Finally, there are many open global questions regarding the regulatory and evolutionary aspects of this intriguing phenomenon, and even the general notion of recoding being adaptively utilized to diversify the proteome is not fully accepted. We look forward to future computational and experimental advancements, combining global analyses of recoding sites and their properties with detailed characterization of individual sites, in hope for clarifying the above questions as well as opening new exciting research directions.

## References

- Alon S et al (2012) Systematic identification of edited microRNAs in the human brain. *Genome Res* 22(8):1533–1540. <https://doi.org/10.1101/gr.131573.111>
- Alon S et al (2015) The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. *eLife* 4. <https://doi.org/10.7554/eLife.05198>
- An NA et al (2019) Evolutionarily significant A-to-I RNA editing events originated through G-to-A mutations in primates. *Genome Biol* 20(1):24. <https://doi.org/10.1186/s13059-019-1638-y>
- Barak M et al (2020) Purifying selection of long dsRNA is the first line of defense against false activation of innate immunity. *Genome Biol* 21(1):26. <https://doi.org/10.1186/s13059-020-1937-3>
- Basilio C et al (1962) Synthetic polynucleotides and the amino acid code. V. *Proc Natl Acad Sci USA* 48(4):613–616

- Bass BL (2002) RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* 71:817–846. <https://doi.org/10.1146/annurev.biochem.71.110601.135501>
- Bass BL, Weintraub H (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* 55(6):1089–1098. [https://doi.org/10.1016/0092-8674\(88\)90253-X](https://doi.org/10.1016/0092-8674(88)90253-X)
- Bazak L et al (2014a) A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* 24(3):365–376. <https://doi.org/10.1101/gr.164749.113>
- Bazak L, Levanon EY, Eisenberg E (2014b) Genome-wide analysis of Alu editability. *Nucleic Acids Res* 42(11):6876–6884. <https://doi.org/10.1093/nar/gku414>
- Behm M et al (2017) Accumulation of nuclear ADAR2 regulates adenosine-to-inosine RNA editing during neuronal development. *J Cell Sci* 130(4):745–753. <https://doi.org/10.1242/jcs.200055>
- Bhalla T et al (2004) Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat Struct Mol Biol* 11(10):950–956. <https://doi.org/10.1038/nsm825>
- Buchumenski I et al (2017) Dynamic hyper-editing underlies temperature adaptation in *Drosophila*. Li JB (eds). *PLOS Genet* 13(7):e1006931. <https://doi.org/10.1371/journal.pgen.1006931>
- Burns CM et al (1997) Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 387(6630):303–308. <https://doi.org/10.1038/387303a0>
- Chalk AM et al (2019) The majority of A-to-I RNA editing is not required for mammalian homeostasis. *Genome Biol* 20(1):268. <https://doi.org/10.1186/s13059-019-1873-2>
- Chen L et al (2013) Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat Med* 19(2):209–216. <https://doi.org/10.1038/nm.3043>
- Chung H et al (2018) Human ADAR1 prevents endogenous RNA from triggering translational shutdown. *Cell* 172(4):811–824.e14. <https://doi.org/10.1016/j.cell.2017.12.038>
- Colina C et al (2010) Regulation of Na<sup>+</sup>/K<sup>+</sup> ATPase transport velocity by RNA editing. *PLoS Biol* 8(11). <https://doi.org/10.1371/journal.pbio.1000540>
- Dagan T et al (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res* 32(Database issue):D489–D492. <https://doi.org/10.1093/nar/gkh132>
- Daniel C et al (2011) Adenosine-to-inosine RNA editing affects trafficking of the gamma-aminobutyric acid type A (GABA(A)) receptor. *J Biol Chem* 286(3):2031–2040. <https://doi.org/10.1074/jbc.M110.130096>
- Daniel C et al (2012) A distant cis acting intronic element induces site-selective RNA editing. *Nucleic Acids Res* 40(19):9876–9886. <https://doi.org/10.1093/nar/gks691>
- Daniel C et al (2014) Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biol* 15(2):R28. <https://doi.org/10.1186/gb-2014-15-2-r28>
- Daniel C et al (2017) Editing inducer elements increases A-to-I editing efficiency in the mammalian transcriptome. *Genome Biol* 18(1):195. <https://doi.org/10.1186/s13059-017-1324-x>
- Diroma MA et al (2017) Elucidating the editome: bioinformatics approaches for RNA editing detection. *Briefings Bioinf*. <https://doi.org/10.1093/bib/bbx129>
- Duan Y et al (2017) Adaptation of A-to-I RNA editing in *Drosophila*. Zhang J (ed). *PLOS Genet* 13(3):e1006648. <https://doi.org/10.1371/journal.pgen.1006648>
- Egebjerg J, Heinemann SF (1993) Ca<sup>2+</sup> permeability of unedited and edited versions of the kainate selective glutamate receptor GluR6. *Proc Natl Acad Sci USA* 90(2):755–759
- Eisenberg E (2012) Bioinformatic approaches for identification of A-to-I editing sites. *Curr Top Microbiol Immunol* 353(1):145–162. [https://doi.org/10.1007/82\\_2011\\_147](https://doi.org/10.1007/82_2011_147)
- Eisenberg E, Levanon EY (2018) A-to-I RNA editing—immune protector and transcriptome diversifier. *Nat Rev Genet* 19(8):473–490. <https://doi.org/10.1038/s41576-018-0006-1>
- Eisenberg E, Li JB, Levanon EY (2010) Sequence based identification of RNA editing sites. *RNA Biol* 7(2):248–252. <https://doi.org/10.4161/rna.7.2.11565>
- Fritz J et al (2009) RNA-regulated interaction of transportin-1 and exportin-5 with the double-stranded RNA-binding domain regulates nucleocytoplasmic shuttling of ADAR1. *Mol Cell Biol* 29(6):1487–1497. <https://doi.org/10.1128/MCB.01519-08>
- Gallo A et al (2017) ADAR RNA editing in human disease; more to it than meets the I. *Hum Genet* 136(9):1265–1278. <https://doi.org/10.1007/s00439-017-1837-0>

- Gal-Mark N et al (2017) Abnormalities in A-to-I RNA editing patterns in CNS injuries correlate with dynamic changes in cell type composition. *Sci Rep* 7:43421. <https://doi.org/10.1038/srep43421>
- Garncarz W et al (2013) A high-throughput screen to identify enhancers of ADAR-mediated RNA-editing. *RNA Biol* 10(2):192–204. <https://doi.org/10.4161/rna.23208>
- Garrett SC, Rosenthal JJC (2012a) A role for A-to-I RNA editing in temperature adaptation. *Physiology* 27(6):362–369. <https://doi.org/10.1152/physiol.00029.2012>
- Garrett S, Rosenthal JJC (2012b) RNA editing underlies temperature adaptation in K<sup>+</sup> channels from polar octopuses. *Science* 335(6070):848–851. <https://doi.org/10.1126/science.1212795>
- George CX et al (2016) Editing of cellular self-RNAs by adenosine deaminase ADAR1 suppresses innate immune stress responses. *J Biol Chem* 291(12):6158–6168. <https://doi.org/10.1074/jbc.M115.709014>
- Gerber A et al (1998) Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. *EMBO J* 17(16):4780–4789. <https://doi.org/10.1093/emboj/17.16.4780>
- Gommans WM, Mullen SP, Maas S (2009) RNA editing: a driving force for adaptive evolution? *BioEssays* 31(10):1137–1145. <https://doi.org/10.1002/bies.200900045>
- Greenberger S et al (2010) Consistent levels of A-to-I RNA editing across individuals in coding sequences and non-conserved Alu repeats. *BMC Genomics* 11(1):608. <https://doi.org/10.1186/1471-2164-11-608>
- Grice LF, Degan BM (2015) The origin of the ADAR gene family and animal RNA editing. *BMC Evol Biol* 15(1):4. <https://doi.org/10.1186/s12862-015-0279-3>
- Hartner JC et al (2009) ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. *Nat Immunol* 10(1):109–115. <https://doi.org/10.1038/ni.1680>
- Higuchi M et al (1993) RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell* 75(7):1361–1370. [https://doi.org/10.1016/0092-8674\(93\)90622-W](https://doi.org/10.1016/0092-8674(93)90622-W)
- Higuchi M et al (2000) Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* 406(6791):78–81. <https://doi.org/10.1038/35017558>
- Hoopengardner B et al (2003) Nervous system targets of RNA editing identified by comparative genomics. *Science (New York, N.Y.)* 301(2003):832–836. <https://doi.org/10.1126/science.1086763>
- Horsch M et al (2011) Requirement of the RNA-editing enzyme ADAR2 for normal physiology in mice. *J Biol Chem* 286(21):18614–18622. <https://doi.org/10.1074/jbc.M110.200881>
- Jain M et al (2018) RNA editing of Filamin A pre-mRNA regulates vascular contraction and diastolic blood pressure. *EMBO J* 37(19). <https://doi.org/10.15252/embj.201694813>
- Jiang D, Zhang J (2019) The preponderance of nonsynonymous A-to-I RNA editing in coleoids is nonadaptive. *Nat Commun* 10(1). <https://doi.org/10.1038/s41467-019-13275-2>
- Kawahara Y et al (2004) Glutamate receptors: RNA editing and death of motor neurons. *Nature* 427(6977):801. <https://doi.org/10.1038/427801a>
- Kawahara Y et al (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* 315(5815):1137–1140. <https://doi.org/10.1126/science.1138050>
- Khermish K et al (2016) Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA (New York, N.Y.)* 22(2):1–13. <https://doi.org/10.1261/rna.054627.115>
- Kleinman CL, Majewski J (2012) Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. *Science* 335(6074):1302; author reply 1302. [335/6074/1302-c \[pii\]. https://doi.org/10.1126/science.1209658](https://doi.org/10.1126/science.1209658)
- Kung SS et al (2001) Q/R RNA editing of the AMPA receptor subunit 2 (GRIA2) transcript evolves no later than the appearance of cartilaginous fishes. *FEBS Lett* 509(2):277–281
- Kwak S, Kawahara Y (2005) Deficient RNA editing of GluR2 and neuronal death in amyotrophic lateral sclerosis. *J Mol Med* 83(2):110–120. <https://doi.org/10.1007/s00109-004-0599-z>
- Levanon EY et al (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* 22(8):1001–1005. <https://doi.org/10.1038/nbt996>

- Levanon EY et al (2005) Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res* 33(4):1162–1168. <https://doi.org/10.1093/nar/gki239>
- Levanon EY, Eisenberg E (2006) Algorithmic approaches for identification of RNA editing sites. *Briefings Funct Genomics Proteomics* 5(1):43–45. <https://doi.org/10.1093/bfgp/ell014>
- Lev-Maor G et al (2007) RNA-editing-mediated exon evolution. *Genome Biol* 8(2):R29. <https://doi.org/10.1186/gb-2007-8-2-r29>
- Li I-C et al (2014a) Zebrafish Adar2 edits the Q/R site of AMPA receptor subunit *gria2 $\alpha$*  transcript to ensure normal development of nervous system and cranial neural crest cells. Sabaawy HE (ed). *PLoS One* 9(5):e97133. <https://doi.org/10.1371/journal.pone.0097133>
- Li Q et al (2014b) Caste-specific RNA editomes in the leaf-cutting ant *Acromyrmex echinator*. *Nat Commun* 5:4943. <https://doi.org/10.1038/ncomms5943>
- Licht K et al (2019) Inosine induces context-dependent recoding and translational stalling. *Nucleic Acids Res* 47(1):3–14. <https://doi.org/10.1093/nar/gky1163>
- Liddicoat BJ et al (2015) RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* 349:1–9. <https://doi.org/10.1126/science.aac7049>
- Lin W et al (2012) Comment on “Widespread RNA and DNA sequence differences in the human transcriptome.” *Science* 335(6074):1302. <https://doi.org/10.1126/science.1210624>
- Liscovitch-Brauer N et al (2017) Trade-off between transcriptome plasticity and genome evolution in cephalopods. *Cell* 169(2):191–202.e11. <https://doi.org/10.1016/j.cell.2017.03.025>
- Lomeli H et al (1994) Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science (New York, N.Y.)* 266(5191):1709–1713. <https://doi.org/10.1126/science.7992055>
- Lonsdale J et al (2013) The genotype-tissue expression (GTEx) project. *Nat Genet* 45(6):580–585. <https://doi.org/10.1038/ng.2653>
- Maas S et al (2001) Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc Natl Acad Sci USA* 98(25):14687–14692. <https://doi.org/10.1073/pnas.251531398>
- Mannon NM et al (2014) The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell Rep* 9(4):1482–1494. <https://doi.org/10.1016/j.celrep.2014.10.041>
- Marcucci R et al (2011) Pin1 and WWP2 regulate *GluR2* Q/R site RNA editing by ADAR2 with opposing effects. *EMBO J* 30(20):4211–4222. <https://doi.org/10.1038/emboj.2011.303>
- Marion S, Weiner DM, Caron MG (2004) RNA editing induces variation in desensitization and trafficking of 5-hydroxytryptamine 2c receptor isoforms. *J Biol Chem* 279(4):2945–2954. <https://doi.org/10.1074/jbc.M308742200>
- Miyake K et al (2016) CAPS1 RNA editing promotes dense core vesicle exocytosis. *Cell Rep* 17(8):2004–2014. <https://doi.org/10.1016/j.celrep.2016.10.073>
- Morgantini C et al (2019) Liver macrophages regulate systemic metabolism through non-inflammatory factors. *Nat Metab* 1(4):445–459. <https://doi.org/10.1038/s42255-019-0044-9>
- Neeman Y et al (2006) RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA* 12(10):1802–1809. rna.165106 [pii]. <https://doi.org/10.1261/rna.165106>
- Nishikura K (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* 17(2):83–96. <https://doi.org/10.1038/nrm.2015.4>
- Oakes E et al (2017) Adenosine deaminase that acts on RNA 3 (ADAR3) binding to glutamate receptor subunit B pre-mRNA inhibits RNA editing in glioblastoma. *J Biol Chem* 292(10):4326–4335. <https://doi.org/10.1074/jbc.M117.779868>
- Ohlson J et al (2007) Editing modifies the GABA(A) receptor subunit  $\alpha$ 3. *RNA (New York, N.Y.)* 13(5):698–703. <https://doi.org/10.1261/rna.349107>
- Palladino MJ et al (2000) dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing (in process citation). *RNA* 6(7):1004–1018. <https://doi.org/10.1017/S1355838200000248>
- Patterson JB, Samuel CE (1995) Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase 15(10):5376–5388

- Patton DE, Silva T, Bezanilla F (1997) RNA editing generates a diverse array of transcripts encoding squid Kv2 K<sup>+</sup> channels with altered functional properties. *Neuron*. [https://doi.org/10.1016/S0896-6273\(00\)80383-9](https://doi.org/10.1016/S0896-6273(00)80383-9)
- Pestal K et al (2015) Isoforms of RNA-editing enzyme ADAR1 independently control nucleic acid sensor MDA5-driven autoimmunity and multi-organ development. *Immunity* 43(5):933–944. <https://doi.org/10.1016/j.immuni.2015.11.001>
- Picardi E et al (2015) Profiling RNA editing in human tissues: towards the inosinome Atlas. *Sci Rep* 5:14941. <https://doi.org/10.1038/srep14941>
- Picardi E et al (2017a) REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res* 45(D1):D750–D757. <https://doi.org/10.1093/nar/gkw767>
- Picardi E, Horner DS, Pesole G (2017b) Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA* 23(6):860–865. <https://doi.org/10.1261/rna.058271.116>
- Pickrell JK, Gilad Y, Pritchard JK (2012) Comment on “Widespread RNA and DNA sequence differences in the human transcriptome.” *Science* 335(6074):1302. <https://doi.org/10.1126/science.1210484>
- Pinto Y, Cohen HY, Levanon EY (2014) Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome Biol* 15(1):R5. <https://doi.org/10.1186/gb-2014-15-1-r5>
- Pinto Y et al (2017) Human cancer tissues exhibit reduced A-to-I editing of miRNAs coupled with elevated editing of their targets. *Nucleic Acids Res*. <https://doi.org/10.1093/nar/gkx1176>
- Piskol R et al (2013) Lack of evidence for existence of noncanonical RNA editing. *Nat Biotechnol* 31(1):19–20. <https://doi.org/10.1038/nbt.2472>
- Porath HT et al (2017a) A-to-I RNA editing in the earliest-diverging eumetazoan phyla. *Mol Biol Evol*. <https://doi.org/10.1093/molbev/msx125>
- Porath HT et al (2017b) Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. *Genome Biol* 18(1):185. <https://doi.org/10.1186/s13059-017-1315-y>
- Porath HT et al (2019) RNA editing is abundant and correlates with task performance in a social bumblebee. *Nat Commun* 10(1):1605. <https://doi.org/10.1038/s41467-019-09543-w>
- Pozo P, Hoopengardner B (2012) Identification and characterization of two novel RNA editing sites in *grin1b* transcripts of embryonic *Danio rerio*. *Neural Plast* 2012:1–7. <https://doi.org/10.1155/2012/173728>
- Ramaswami G, Li JB (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* 42(D1):D109–D113. <https://doi.org/10.1093/nar/gkt996>
- Ramaswami G, Li JB (2016) Identification of human RNA editing sites: a historical perspective. *Methods* 107:42–47. <https://doi.org/10.1016/j.jmeth.2016.05.011>
- Ramaswami G et al (2015) Genetic mapping uncovers cis-regulatory landscape of RNA editing. *Nat Commun* 6:8194. <https://doi.org/10.1038/ncomms9194>
- Rebagliati MR, Melton DA (1987) Antisense RNA injections in fertilized frog eggs reveal an RNA duplex unwinding activity. *Cell* 48(4):599–605
- Reenan RA (2005) Molecular determinants and guided evolution of species-specific RNA editing. *Nature* 434(7031):409–413. <https://doi.org/10.1038/nature03364>
- Reich DP, Bass BL (2019) Mapping the dsRNA world. *Cold Spring Harbor Perspect Biol* 11(3):a035352. <https://doi.org/10.1101/cshperspect.a035352>
- Rieder LE et al (2013) Tertiary structural elements determine the extent and specificity of messenger RNA editing. *Nat Commun* 4:2232. <https://doi.org/10.1038/ncomms3232>
- Rieder LE et al (2015) Dynamic response of RNA editing to temperature in *Drosophila*. *BMC Biol* 13(1):1. <https://doi.org/10.1186/s12915-014-0111-3>
- Riemondy KA et al (2018) Dynamic temperature-sensitive A-to-I RNA editing in the brain of a heterothermic mammal during hibernation. *RNA (New York, N.Y.)* 24(11):1481–1495. <https://doi.org/10.1261/rna.066522.118>
- Robinson JE et al (2016) ADAR-mediated RNA editing suppresses sleep by acting as a brake on glutamatergic synaptic plasticity. *Nat Commun* 7. <https://doi.org/10.1038/ncomms10512>

- Rosenthal JJC, Bezanilla F (2002) Extensive editing of mRNAs for the squid delayed rectifier K<sup>+</sup> channel regulates subunit tetramerization. *Neuron* 34(5):743–757. [https://doi.org/10.1016/S0896-6273\(02\)00701-8](https://doi.org/10.1016/S0896-6273(02)00701-8)
- Roth SH, Levanon EY, Eisenberg E (2019) Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. *Nat Methods* 16(11):1131–1138. <https://doi.org/10.1038/s41592-019-0610-9>
- Rueter SM, Dawson TR, Emeson RB (1999) Regulation of alternative splicing by RNA editing. *Nature* 399(6731):75–80. <https://doi.org/10.1038/19992>
- Sailer A et al (1999) Generation and analysis of GluR5 (Q636R) kainate receptor mutant mice. *J Neurosci Off J Soc Neurosci* 19(20):8757–8764
- Sapiro AL et al (2015) Cis regulatory effects on A-to-I RNA editing in related *Drosophila* species. *Cell Rep* 11(5):697–703. <https://doi.org/10.1016/j.celrep.2015.04.005>
- Savva YA et al (2012) Auto-regulatory RNA editing fine-tunes mRNA re-coding and complex behaviour in *Drosophila*. *Nat Commun* 3:790. <https://doi.org/10.1038/ncomms1789>
- Schneider WM, Chevillotte MD, Rice CM (2014) Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol* 32(1):513–545. <https://doi.org/10.1146/annurev-immunol-032713-120231>
- Schrider DR, Gout J-F, Hahn MW (2011) Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One* 6(10):e25842. <https://doi.org/10.1371/journal.pone.0025842>
- Seeburg PH, Hartner J (2003) Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Curr Opin Neurobiol* 13(3):279–283
- Shamay-Ramot A et al (2015) Fmrp interacts with Adar and regulates RNA editing, synaptic density and locomotor activity in zebrafish. *PLoS Genet* 11(12):e1005702. <https://doi.org/10.1371/journal.pgen.1005702>
- Sie CP, Maas S (2009) Conserved recoding RNA editing of vertebrate C1q-related factor C1QL1. *FEBS Lett* 583(7):1171–1174. <https://doi.org/10.1016/j.febslet.2009.02.044>
- Sommer B et al (1991) RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* 67(1):11–19. 0092-8674(91)90568-J [pii]
- Sorek R et al (2004) Minimal conditions for exonization of intronic sequences: 5' splice site formation in Alu exons. *Mol Cell* 14(2):221–231. [https://doi.org/10.1016/S1097-2765\(04\)00181-9](https://doi.org/10.1016/S1097-2765(04)00181-9)
- Tan MH et al (2017) Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550(7675):249–254. <https://doi.org/10.1038/nature24041>
- Terajima H et al (2016) ADARB1 catalyzes circadian A-to-I editing and regulates RNA rhythm. *Nat Genet* 49(1):146–151. <https://doi.org/10.1038/ng.3731>
- Tian N et al (2008) A-to-I editing sites are a genomically encoded G: implications for the evolutionary significance and identification of novel editing sites. *RNA (New York, N.Y.)* 14(2):211–216. <https://doi.org/10.1261/rna.797108>
- Vesely C et al (2012) Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. *Genome Res* 22(8):1468–1476
- Wahlstedt H et al (2009) Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res*. <https://doi.org/10.1101/gr.089409.108>
- Wang Q et al (2000) Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* 290(5497):1765–1768. <https://doi.org/10.1126/science.290.5497.1765>
- Wang Y et al (2017) Systematic characterization of A-to-I RNA editing hotspots in microRNAs across human cancers. *Genome Res* 27(7):1112–1125. <https://doi.org/10.1101/gr.219741.116>
- Wu J, Chen ZJ (2014) Innate immune sensing and signaling of cytosolic nucleic acids. *Annu Rev Immunol* 32:461–488. <https://doi.org/10.1146/annurev-immunol-032713-120156>
- Xu G, Zhang J (2014) Human coding RNA editing is generally nonadaptive. *Proc Natl Acad Sci USA* 111(10):3769–3774. <https://doi.org/10.1073/pnas.1321745111>
- Yablonovitch AL et al (2017a) Regulation of gene expression and RNA editing in *Drosophila* adapting to divergent microclimates. *Nat Commun* 8(1):1570. <https://doi.org/10.1038/s41467-017-01658-2>

- Yablonovitch AL et al (2017b) The evolution and adaptation of A-to-I RNA editing. Zhang J (ed). PLOS Genet 13(11):e1007064. <https://doi.org/10.1371/journal.pgen.1007064>
- Yeo J et al (2010) RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. Proc Natl Acad Sci USA 107(48):20715–20719. <https://doi.org/10.1073/pnas.1009231107>
- Yu Y et al (2016) The landscape of A-to-I RNA editome is shaped by both positive and purifying selection. Schierup MH (ed). PLOS Genet 12(7):e1006191. <https://doi.org/10.1371/journal.pgen.1006191>
- Zaidan H et al (2018) A-to-I RNA editing in the rat brain is age-dependent, region-specific and sensitive to environmental stress across generations. BMC Genomics 19(1):28. <https://doi.org/10.1186/s12864-017-4409-8>
- Zhang S-J et al (2014) Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. Mol Biol Evol 31(5):1309–1324. <https://doi.org/10.1093/molbev/msu084>
- Zhang R et al (2017) Evolutionary analysis reveals regulatory and functional landscape of coding and non-coding RNA editing. Zhang J (ed). PLOS Genet 13(2):e1006563. <https://doi.org/10.1371/journal.pgen.1006563>

# Chapter 5

## Most Successful Mammals in the Making: A Review of the Paleocene Glires



Łucja Fostowicz-Frelik

**Abstract** Glires, the most speciose clade of placental mammals nowadays includes the conservative, uniformly shaped lagomorphs and widely diversified rodents. Both groups are recognized in the fossil record since the early Paleogene, but Rodentia appeared slightly earlier (the late Paleocene) than lagomorphs of modern aspect (the early Eocene). The earliest Glires currently recognized come from Asia (East China), where they were scarce and relatively poorly diversified. They are basal taxa of neither rodent nor lagomorph clear affiliations, although probably ancestral for both. In contrast, the earliest record of North American Glires consists of scarce Rodentiaformes (known also from the late Paleocene of Asia) and the primitive basal rodents Ischyromyidae, the latter a widely diversified and abundant group. Ischyromyid rodents differ from other basal Glires lineages in generally large size, approximately that of ground squirrels or marmots. Ischyromyids are an important example of “the Paleocene Paradox” in the fossil record, which is a discrepancy between the fact that the earliest fossil rodents are known from North America while it is widely presumed that the Rodentia originated in Asia from earlier gliroid mammals. Here, I provide a brief review of the early diversification of Glires both in Asia and North America, and discuss the earliest morphologies in this group. Two related important points in the beginning of the evolutionary history of Glires are the skull structure as a whole and the dentition.

---

L. Fostowicz-Frelik (✉)

Key Laboratory of Mammal Evolution and Human Origins, Institute of Vertebrate Paleontology and Anthropology, Chinese Academy of Sciences, Beijing 100044, People’s Republic of China  
e-mail: [lucja\\_fostowicz@yahoo.com](mailto:lucja_fostowicz@yahoo.com)

CAS Center for Excellence in Life and Palaeoenvironment, Beijing 100044,  
People’s Republic of China

Institute of Paleobiology, Polish Academy of Sciences, 00-818 Warsaw, Poland

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_5](https://doi.org/10.1007/978-3-030-57246-4_5)



## 5.1 Introduction

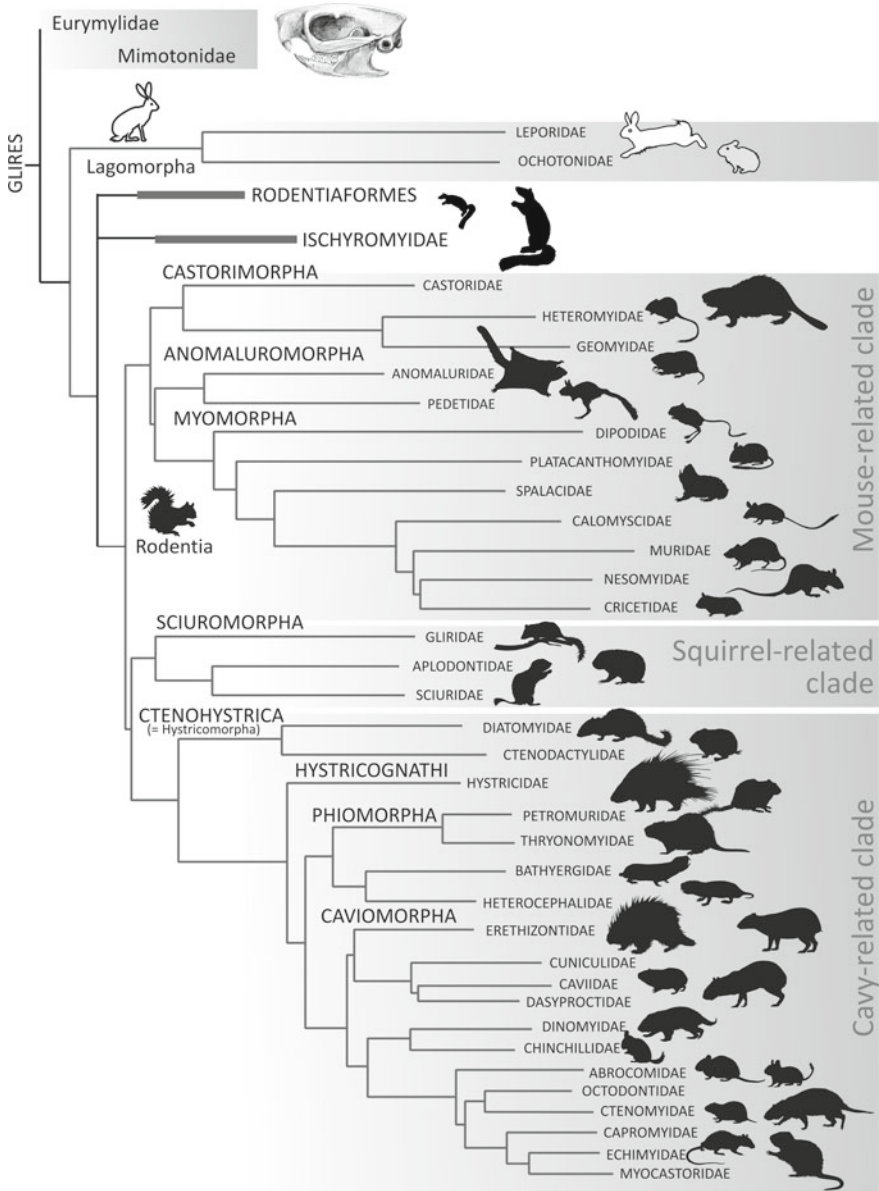
Glires (rodents and lagomorphs), the counterpart to Euarchonta (primates, tree shrews and colugos) within the Euarchontoglires clade (Fig. 5.1), which is one of four major clades of modern placental mammals (e.g., Murphy et al. 2001; Asher et al. 2009), are the most successful and speciose group of living mammals (Wilson et al. 2016). The extant Glires consist of two clades: the lagomorphs, grouping rabbits, hares, and pikas, and rodents, including a wide variety of animals divided into three major clades: Ctenohystrica (e.g., gundis, porcupines, and guinea pigs), Myomorpha (e.g., hamsters, mice, rats, and beavers), and Sciuromorpha (squirrels in a broad sense).

The lagomorphs represent a conservative, basal branch of Glires (see e.g., Asher et al. 2019). Including ca. 95 living species, this clade is poorly diversified in comparison with rodents, the clade including over 2400 living species (Wilson et al. 2016). Extant rodents are widely diversified (Fig. 5.1), representing almost all locomotor adaptations of terrestrial mammals and inhabiting most of the accessible terrestrial habitats worldwide (Wilson et al. 2016). Also, rodents as a whole express most of known dietary preferences. Most of the species is herbivorous, including frugivorous and granivorous forms, and interestingly, some rodents are fully carnivorous. They are either exclusively insectivorous (e.g., *Selevinia*) or piscivorous and “crab-eaters” (the Ichthyomyidae; see Voss 1988). Also, taking into account their body mass, modern Glires are quite diverse in size. Rodents exhibit wide size differences, ranging from the Baluchistan pygmy jerboas (*Salpingotulus michaelis*) weighting only a few grams to the biggest extant rodents, capybaras (60 kg). The fossil record is even more astounding, including the giant beavers, and the Pliocene dinomyid *Josephoartigasia*, the largest rodent ever known that reached about 1000 kg (Rinderknecht and Blanco 2008).

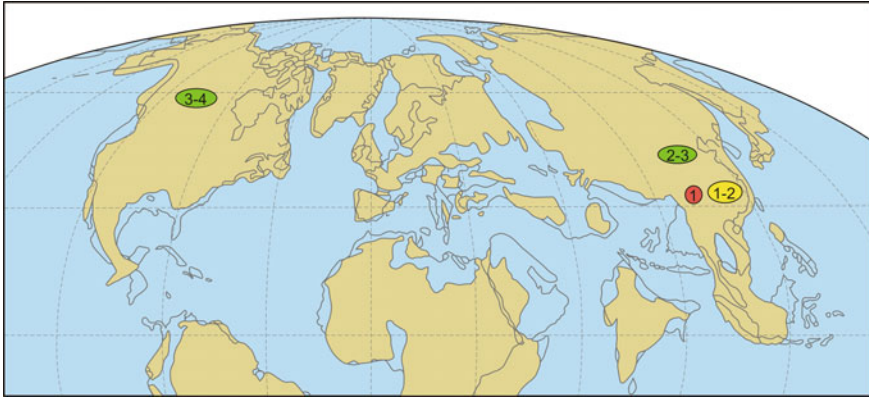
Glires are an archaic group known already from the Early Paleocene (Fig. 5.2). Their diversity in form and size was growing constantly since their inception, showing several significant accelerations, especially at the beginning of the Oligocene and Miocene, and during the Mio-Pliocene interval. But what can be said of the earliest forms? Does the Paleocene fossil record of Glires shows this considerable morphological potential or was it rather conservative, only gradually gaining momentum? This chapter analyzes the morphological variability of earliest known Glires constituting the Paleocene fossil record.

## 5.2 Origins of Glires in Asia

There is a general consensus that Glires originated in Asia (Li 1977; Li and Ting 1993; Meng et al. 2003; Meng and Wyss 2005). The first remains referred to this group, the basal members of Glires come from the early Paleocene of the Wanghudun Formation in Qianshan (Anhui, China). These earliest Glires cannot be assigned



**Fig. 5.1** Phylogenetic scheme of Glires relationships. Phylogeny of extant clades based on maximum likelihood analysis of 31 genes and 39,000 nucleotide base pairs (see Wilson et al. 2016); the position of stem rodent groups and basal Glires follows Asher et al. (2005). Animals silhouettes not to scale



**Fig. 5.2** Worldwide distribution of the Paleocene Glires. 1, Mimotonidae; 2, Eurymylidae; 3, Rodentiaformes (Alagomyidae); 4, Ischyromyidae (stem Rodentia). Color code: red, early Paleocene; yellow, middle Paleocene; green, late Paleocene. The map shows continent arrangement at the Paleocene/Eocene transition, based on Smith et al. (1994)

neither to Rodentia nor Lagomorpha, although they show features considered ancestral to both groups (Sych 1971; Averianov 1994) and already express the diversity due to exact count and shape of the incisors, one of the most often used factors to distinguish these two groups. The number of the upper incisors which may be one pair (Simplicidentata) or two (Duplicidentata) is the foundation for a very basic distinction between different Glires (Asher et al. 2005). Generally, Duplicidentata are linked with lagomorpha lineage (and formally include the extant lagomorphs), while Simplicidentata are thought to be inclusive of extant rodents and their ancestral stock (Averianov 1994). This may be, however, at least partly expression of evolutionary mosaicity or frequent parallelisms observed within the Glires (see Fostowicz-Frelik 2017).

The notably scarce first fossil record of the early to middle Paleocene Glires is limited to central China and consists of three genera with five species (Table 5.1), represented mostly by the type material (Li 1977; Li et al. 2016; Wang et al. 2016).

The situation changes during the Late Paleocene, when the first wave of radiation within Glires occurs. In Asia at that time, ten more basal Glires appear (Table 5.1), but their occurrence area is shifted north, to the Mongolian Plateau (Mongolia and Nei Mongol Autonomous Region of China). Furthermore, in the Late Paleocene, the first Rodentiaformes appear, both in the Mongolian Plateau (Dashzeveg and Russell 1988; Meng and Wyss 2001) and North America (Dawson and Beard 1996; Flynn 2008), where they co-occur with the first true rodents of the Ischyromyidae group. These groups differ already markedly in size, dental structure, and skeleton morphology laying the foundation for further diversification, although the common ancestral pattern still can be traced in their morphology.

**Table 5.1** Worldwide record of the Paleocene Glires

Species		Material available	Age	Formation/Horizon and locality	Comments and references
<b>Basal Glires</b>					
<i>Amar aleator</i>		Dental (holotype only)	Late Paleocene	Zhigden Member of the Naran Bulak Formation: Tsagan Kushu Quarry 3 (Mongolia)	Dashzeveg and Russell (1988)
<i>Eomylytus bayanulanensis</i>		Teeth, mandibular, and maxillar fragments	Late Paleocene	Bayan Ulan Formation: Bayan Ulan, Nei Mongol (China)	Meng et al. (2005)
<i>Eomylytus borealis</i>		Teeth, mandibular	Late Paleocene	Bayan Ulan Formation: Bayan Ulan, Nei Mongol (China)	Chow and Qi (1978), Dashzeveg and Russell (1988), Meng et al. (1998)
<i>Eomylytus zhigdenensis</i>		Teeth, mandibular, and maxillar fragments	Late Paleocene to earliest Eocene (Bumbarian ALMA)	Zhigden Member: Tsagan Kushu Quarry 3; Naran Bulak Quarry 1 (Nemegt Basin, Mongolia) Naran Member: Naran Bulak Quarry 1 (Nemegt Basin, Mongolia) Bottom of Bumban Member: Tsagan Kushu Quarry 1, (Nemegt Basin, Mongolia)	Dashzeveg and Russell (1988), Dashzeveg et al. (1998)
<i>Eurymylus laticeps</i>		Dental, incomplete cranial	Late Paleocene	Gashato Formation (Khashat Member 1): Gashato (Ulan Nur Basin, Mongolia)	Matthew and Granger (1925), Sych (1971), Dashzeveg and Russell (1988)
<i>Hanomys malcolmi</i>		Dental, incomplete cranial	Late Paleocene	Bamiao, Xiaochuan town (Hubei, China)	Huang et al. (2004)

(continued)

Table 5.1 (continued)

Species	Material available	Age	Formation/Horizon and locality	Comments and references
<i>Heomys orientalis</i>	Dental, incomplete cranial	Early to middle Paleocene	The lower part of the Upper Member of the Wanghudun Formation: Zhangjiawu Southwest, Qianshan (Anhui, China) Upper Member of Doumu Formation: Yangxiaowu, Qianshan (Anhui, China)	Li (1977), Wang et al. (2016)
<i>Khaychima elongata</i>	Dental, mandibular	Late Paleocene	Naran Member: Khaychin-Ula 1 (Bugin Tsav Basin, Mongolia)	Dashzeveg and Russell (1988)
<i>Mimotona lii</i>	Dental, mandibular (holotype only)	Early Paleocene	Lower part of the Upper Member of Wanghudun Formation: Zhangjiawu south, Qianshan (Anhui, China)	Li (1977), Dashzeveg and Russell (1988)
<i>Mimotona robusta</i>	Dental, mandibular (holotype only)	Middle Paleocene	Lower Member of the Doumu Formation: Hanhuawu south, Qianshan (Anhui, China)	Li (1977)
<i>Mimotona wana</i>	Dental, incomplete cranial? Postcranial (calcaneus)	(1) Early to (2) middle Paleocene	(1) Lower part of the Upper Member of Wanghudun Formation: Shangxialou, Qianshan (Anhui, China) (2) Upper Member of the Doumu Formation: Yangxiaowu (Anhui, China)	Li (1977), Li and Ting (1985, 1993)

(continued)

Table 5.1 (continued)

Species	Material available	Age	Formation/Horizon and locality	Comments and references
<i>Mina hui</i>	Dental: upper incisors and upper molars (holotype only)	Early middle Paleocene	Upper part of the Upper Member of Wanghudun Formation: Fujiaoshanzui, Qianshan (Anhui, China)	Li et al. (2016)
<i>Palaeomytus lii</i>	Dental, mandibular	Late Paleocene	Bayan Ulan Formation: Bayan Ulan (Nei Mongol, China)	Meng et al. (2005)
<i>Sinomylus zhaiti</i>	Dental, incomplete cranial	Late Paleocene	Tujinshan Formation: Jiashan Xian (Anhui, China)	McKenna and Meng (2001)
<i>Taizimylus tongi</i>	Dental, incomplete cranial	Late Paleocene	Xinjiang, China	Mao et al. (2017)
Rodentiaformes				
<i>Alagomys russelli</i>	Dental	Late Paleocene (late Clarkforkian)	Fort Union Formation: Big Multi Quarry, Washakie Basin (Wyoming, USA)	Dawson and Beard (1996)
<i>Neimengomys qii</i>	Dental	Late Paleocene	Gashatan: Subeng, Erlian Basin (Nei Mongol, China)	Meng et al. (2007)
<i>Tribosphenomys minutus</i>	Dental, incomplete cranial, postcranial	Late Paleocene	Lowest part of the Bayan Ulan beds Nomogen Formation: Bayan Ulan and Subeng, Erlian Basin (Nei Mongol, China)	Meng et al. (1994, 1998, 2007), Meng and Wyss (2001)
<i>Tribosphenomys secundus</i>	Dental	Late Paleocene	(1) Zhigden Member, Naran Bulak Formation: Tsagan Kushu (Nemegt Basin, Mongolia) (2) Gashatan: Subeng, Erlian Basin (Nei Mongol, China)	Lopatini and Averianov (2004a), Meng et al. (2007)

(continued)

Table 5.1 (continued)

Species	Material available	Age	Formation/Horizon and locality	Comments and references
<i>Tribosphenomys tertius</i>	Dental	Late Paleocene	Zhigden Member, Naran Bulak Formation: Tsagan Kushu (Nemegt Basin, Mongolia)	Lopatin and Averianov (2004b)
Basal Rodentia				
<i>Acritoparamys atavus</i> <i>A. atwateri</i> <i>A. francesi</i>	Dental	Late Paleocene (Clarkforkian 2-3 NALMA)	Fort Union and Willwood Formations: Bighorn, Clark's Fork, and Washakie basins (Wyoming, USA)	Rose (1981), Ivy (1990), Anderson (2008)
<i>Framimys amherstensis</i>	Dental, cranial	Late Paleocene (Clarkforkian 2-3 NALMA)	Fort Union and Willwood Formations: Bighorn, Clark's Fork basins (Wyoming, USA)	Rose (1981), Ivy (1990), Anderson (2008)
<i>Microparamys cheradius</i> <i>M. minutus</i>	Dental, cranial	Late Paleocene (Clarkforkian 2-3 NALMA)	Fort Union and Willwood Formations: Bighorn, Clark's Fork, and Washakie basins (Wyoming, USA)	Ivy (1990), Anderson (2008)
<i>Paramys adamus</i>	Dental, mandibular	Late Paleocene (Clarkforkian 2 NALMA)	Fort Union Formation: Big Multi Quarry Local Fauna, Washakie Basin (Wyoming, USA)	Dawson and Beard (1996)
<i>Paramys taurus</i>	Dental	Late Paleocene (Clarkforkian 2-3 NALMA)	Fort Union and Willwood Formations: Bighorn, Clark's Fork basins (Wyoming, USA)	Rose (1981), Ivy (1990), Anderson (2008)

### 5.3 Basal Glires: Eurymylidae and Mimotonidae

The most primitive forms assigned to Glires regarded as basal or stem groups are included into two families: Eurymylidae and Mimotonidae. They form a paraphyletic cluster nested at the root of the phylogenetic tree of Glires (Meng and Wyss 2001; Meng et al. 2003; Asher et al. 2005). Historically, this group was known as “Mixodontia” (Sych 1971). The name was meant to reflect their dental morphology, which shows a mixture of lagomorph and rodent characteristics, summing up mostly to the formation of the incisor segment (two or one pair), lack of canines and anteriormost premolars, and a nascent unilateral hypsodonty of the molars (and existing premolars), which in a more developed state is characteristic of earliest Lagomorpha (Li 1977; Li et al. 2007).

As the group is paraphyletic (both as “Mixodontia,” as well as each family separately), it should be regarded more in terms of a morphological grade, which includes animals with a similar lifestyle and diet preferences, than in strictly phylogenetic perspective.

The oldest known Glires representatives are mimotonidae, small, and scarce mammals (Li 1977; Li and Ting 1993), regarded as closer to lagomorphs. The Paleocene record of the group consists of only two genera (*Mimotona* and *Mina*) and includes four species (Table 5.1), all known exclusively from the early and middle Paleocene of Qianshan area in Anhui Province, China (Li 1977; Li et al. 2016). Two species of *Mimotona* (*M. lii* and *M. wana*) were recovered from the early Paleocene strata, dated at ca. 62 Ma (Wang et al. 2016), quite close to the K/Pg boundary at 66 Ma. This fact makes them one of the earliest true placentals, which can be linked with crown (extant) groups of Placentalia. Both *Mimotona* and *Mina* have two pairs of upper incisors, the feature which unite them with Lagomorpha, they also have (at least, it is confirmed for *Mimotona*) two pairs of lower incisors, an apparently primitive feature for Glires, implied also for some eurymylids (see Meng et al. 2005).

The Paleocene mimotonids represent different lineage than the Eocene representatives of the group: *Gomphos* and *Mimolagus* (see Bohlin 1951; Asher et al. 2005; Fostowicz-Frelik et al. 2015). A similar size disparity is observed also between the Paleocene and Eocene eurymylids, the latter being much larger and expressing different dental (and cranial) adaptations, indicating certain more directional adaptations (see Meng et al. 2003). In fact, the Paleocene mimotonids show less similarities to the Eocene ones than to Lagomorpha of modern aspect.

The Paleocene is decidedly the acme for Eurymylids, another group of basal Glires (Sych 1971; Dashzeveg and Russell 1988; Meng et al. 2005); with almost 80% of species belonging to this group known solely from this period (Mao et al. 2017). Similarly to early mimotonids, the Paleocene eurymylids were small animals with approximate skull length not exceeding four centimeters (see Sych 1971; Li 1977). Eurymylids are undoubtedly the most primitive Simplicidentata, the Glires group having a single pair of incisors; but in the light of eurymylid paraphyly, it is



obvious that Simplicidentata may be yet another morphological grade in this respect. Therefore, the reduction of the number of incisors may have been parallel in several different basal Glires lineages.

#### 5.4 Paleocene Glires in North America: Alagomyidae and Ischyromyidae

Two new groups of Glires appeared during the Late Paleocene, and they are pre-rodent Rodentiaformes Alagomyidae (Meng and Wyss 2001; Wyss and Meng 1996) and Ischyromyidae, the basalmost group of rodents (Table 5.1). Interestingly, alagomyids are known from Asia as well as from North America, whereas the Paleocene ischyromyids are known only from North America.

The phylogenetic status of both groups is uncertain, they are treated either as the primitive Rodentia (Flynn 2008) or as a sister group to the *Paramys*–*Cocomys* clade of true Rodentia (see Dawson and Beard 1996) or each of them is placed as a different stem lineage of Rodentia (understood in modern aspect), outside of the crown rodents (see Meng and Wyss 2005). The phylogenetic status of Rodentiaformes as a pre-rodent group is much less controversial than the position of Ischyromyidae (Wyss and Meng 1996; Meng and Wyss 2001, 2005). The latter group was conventionally classified as rodents and was proposed as ancestral to Sciuridae (Wood 1962), although their morphological divergence from other early rodent groups (in particular of earliest ctenodactyloid hystricognathes Cocomyidae) is striking. Ischyromyids are the only large Glires in the Paleocene, and they exceeded the size of any other Glires known from this period and even most of the early Eocene ctenodactyloids (see dental data in Anderson 2008). Although little is known about the postcranial morphology of the Paleocene representatives, the early Eocene (Wasatchian NALMA) conspecific or congeneric ischyromyid findings can be informative in this respect (Wood 1962; Rose and Chinnery 2004). As the ischyromyids seem rather conservative in morphology, both dentally and skeletally, thus it can be safely assumed that the Paleocene forms did not depart far from the morphotype of the Eocene members. They were rather strongly built, with a slightly elongated body, the powerful limbs, and a long strong tail, resembling somewhat ground squirrels (Rose and Chinnery 2004). On the other hand, Rodentiaformes are very small, showing gracile foot structure with elongated lower tarsal bones and metatarsal foot segment (see Meng and Wyss 2001). Nevertheless, concerning the tarsal morphology, they showed quite different adaptations from those observed in earliest true rodents of the ctenodactyloid group (see Fostowicz-Frelik et al. 2018).

The rodentiaform alagomyids and ischyromyids showed more derived dental adaptations and more pronounced reductions of the premolars compared to basal Glires. *Tribosphenomys* has a very reduced peg-like P3, no p3, and its P4/p4 is molarized, although alagomyids differ in the degree of molarization of the premolar loci (see Dawson and Beard 1996 and references therein). Alagomyids also frequently

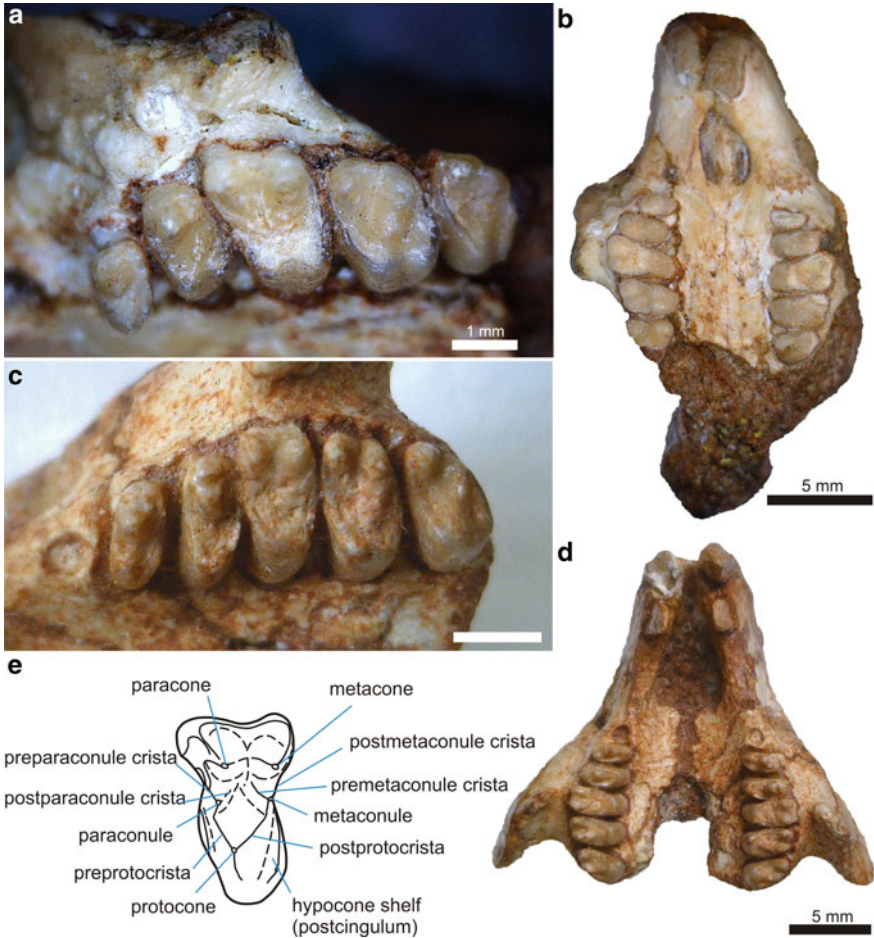
retain deciduous fourth premolars (Meng and Wyss 2001) and thus delay the eruption of the permanent teeth, which may be the evolutionary prerequisite to the permanent loss of P4/p4 (the loci frequently missing in modern rodent groups). Similarly, ischyromyids still retain a reduced P3, and they tend to molarize the P4/p4 (Anderson 2008).

## 5.5 Dentition

The dental formula of basalmost Glires is modified in comparison with the dental formula of primitive eutherians (see Kielan-Jaworowska et al. 2004). Glires have two pairs of upper and lower incisors maximum and show a progressive loss of the premolars (Meng and Wyss 2005); thus, the generalized dental formula should read I1-2/1-2 C0/0 P0-3/0-2 M 2-3/2-3. The earliest basal Glires, such as *Mimotona*, express the dental formula (I2/2 C 0/0 P3/2 M 3/3) almost identical to that of modern leporids, but the latter have a single pair of lower incisors. It is noteworthy that the anteriormost incisors in modern Glires were established on the basis of developmental studies to be the DI2/di2 loci (see Meng et al. 2003). The posterior pair of the incisors, if exists, represents the pair I3/i3 of the incisors in the placental model and has normally two dental generations. This homology is assumed for all Glires whether modern or extinct (Meng et al. 2003).

The loss of the second pair of incisor is apparently a progressive character, and it appears first in the upper dentition. All eurymylids including *Heomys* (Fig. 5.3a, b) have already only a single pair of the upper incisors (DI2). The upper incisors of eurymylids are large, compressed medio-laterally (Fig. 5.3b, 5.4), looking proportionally larger than the teeth in most rodents (Sych 1971; Li 1977). In the Paleocene mimotonids, these teeth are more delicate (Fig. 5.3d). *Mimotona* displays a groove at the anterior surface of the DI2 (Li 1977; Li and Ting 1985, 1993), and its anterior incisors are compressed anteroposteriorly, showing almost identical morphology to that observed in modern lagomorphs. On the other hand, *Mina* does not have such a groove, and both pairs of its upper incisors are compressed mediolaterally as in eurymylids (Li et al. 2016). The incisor morphology of alagomyids and ischyromyids resembles that of eurymylidae; *Tribosphenomys* shows a rather great curvature of the anterior upper incisor (Meng and Wyss 2001).

The premolar reduction is progressive in Glires, and the earliest members of the group (e.g., *Mimotona*, *Sinomylus*, and *Taizimylus*) have three upper premolars (Li 1977; McKenna and Meng 2001; Mao et al. 2017), the dental pattern which is preserved also in extant Lagomorpha. But the middle Paleocene *Heomys* and the late Paleocene *Eurymylus* and *Eomylus* already lost P2 (Sych 1971; Li 1977; Dashzeveg and Russell 1988). The lower dentition is more uniform. All the basal Glires have retained p3 and p4, the condition observed in modern lagomorphs. Rodents have much more progressive reductions as they primitively have only one premolar (p4/P4), and in most crown groups (especially murids) they lost even this premolar locus. Rodentiaformes display a simple upper P3 and molarized DP4. The

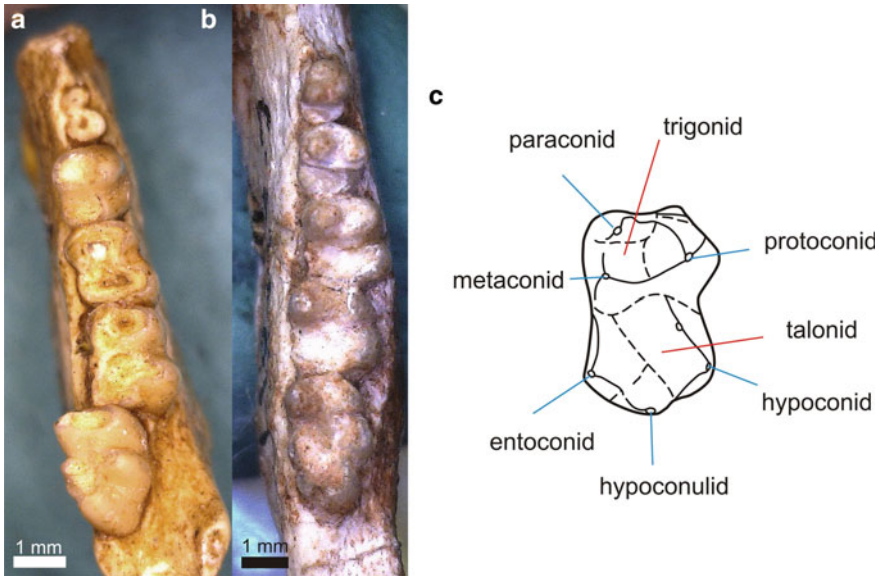


**Fig. 5.3** Upper dentition of the earliest Glires. **a, b** The simplicitate *Heomys orientalis* (IVPP V4321); **c, d**, the duplicitate *Mimotona wana* (IVPP V7416). **a** Right upper tooth row (mirror view); **b** ventral view of the skull; **c** left upper tooth row; **d** ventral view of the skull; **e** explanatory drawing of main dental structures in a tribosphenic molar; modified from Kielan-Jaworowska et al. (2004)

P3 is reduced to a small single-cusped, peg-like tooth (Meng and Wyss 2001). The P4 is apparently delayed in ontogeny and is mostly lacking in known alagomyid material. Thus, the functional posteriormost premolar is a deciduous tooth (DP4).

The rodentiaform lower tooth row includes only one premolar, which is regularly exchanged into a non-molarized permanent tooth (p4). This locus is more simplified than its deciduous predecessor, dp4 (Meng and Wyss 2001).

Ischyromyid rodents display the dental pattern similar to that of the Rodentiiformes. Most of ischyromyids retain P3 and P4/p4; the ultimate premolars are at least partly molarized, although still smaller than the molars (Anderson 2008).



**Fig. 5.4** Lower dentition of the earliest Glires. **a** The simplicidentate *Heomys orientalis* (IVPP V4322; right mandible with p4–m3); **b** the duplicitentate *Mimotona wana* (IVPP V 7416.1; left mandible with p3–m3, mirror view); **c** explanatory drawing of main dental structures in a tribosphenic molar; modified from Kielan-Jaworowska et al. (2004). Note the similarities in both groups

The structural pattern of the occlusal tooth surface in Glires comprising of cusps, crests, and basins adheres to the tribosphenic molar pattern typical of all placentals. Nevertheless, the dentition of modern Glires representatives departed far from the original morphotype. The dental pattern of the basalmost Glires (see *Heomys* and *Mimotona* in Fig. 5.3) closely resembles the typical tribosphenic molar of the Cretaceous eutherian having all basal structures (see Fig. 5.3). The upper cheek teeth are anteroposteriorly compressed and extended mediolaterally. Both genera differ very little in general dental morphology. Three major cusps, the paracone, metacone, and protocone are well developed and recognizable, the metaconules are usually larger than the paraconules, hypocones and hypoconal shelves (postcingula) are well developed, and the trigon basin is relatively large. Compared to the late Cretaceous Zalambdalestidae (see Kielan-Jaworowska et al. 2004), the lingual part of the tooth is better developed in Glires, the hypocone is much stronger and the hypoconal shelf is larger. The lower cheek dentition of *Heomys* and *Mimotona* differs even less from the typical tribosphenic pattern of the late Cretaceous Eutheria (Fig. 5.4; see also Fostowicz-Frelik 2016). The only significant difference is the reduction of the paraconid (Meng et al. 2003). Some eurymylids (e.g., *Eomylus*) have still a small paraconid (or at least a well-developed paracristid), and this cusp is also observed in the deciduous premolars of early Glires, but all mimotonids lack this character in their permanent dentition (the character in common with lagomorphs).

The Rodentiaformes dental pattern also does not depart markedly from the tribosphenic one. Alagomyids show some reductions in the hypocone and hypoconal shelf, which are interpreted as progressive in their lineage (Dawson and Beard 1996).

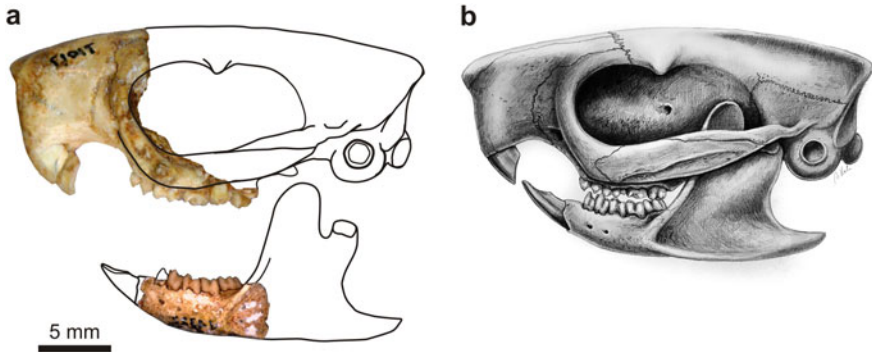
On the other hand, the occlusal dental pattern of ischyromyids shows further modifications typical for rodents. The teeth become square in the outline, the cingulids (the pre- and post-cingulum) become stronger and better developed, while the cusps are joined with more strongly developed cristae, gradually coalescing with them, and the center of the trigonid and talonid forms more extended and deeper basins (see Meng et al. 2005; Anderson 2008).

## 5.6 The Skull

Very little is known about the skull of the Paleocene Glires. Among the stem taxa, the anterior part of the cranium is known for eurymylids: *Eurymylus*, *Hanomys*, *Heomys*, *Sinomylus*, and *Taizimylus* (Sych 1971; Li 1977; McKenna and Meng 2001; Huang et al. 2004; Mao et al. 2017), and the mimotonid *Mimotona wana* (Li and Ting 1993). Similarly, only very small fragment of the anterior part of the skull is known for the Rodentiaformes *Tribosphenomys minutus* from the late Paleocene Subeng locality in Nei Mongol, China (Meng and Wyss 2001; Meng et al. 2007). In case of the Paleocene ischyromyids, the cranial material is not known, although the early Eocene (Wasatchian) findings representing the same genera (and species) include cranial material (Rose and Chinnery 2004).

In all cases of the Paleocene stem Glires skulls, the morphology of the substantial part of the muzzle (with the nasals and premaxillae) and the partially preserved frontals, maxillae, and the anterior roots of the zygomatic arches is known (see e.g., Fig. 5.5). In most cases the hard palate (Fig. 5.3b, d), consisting of the maxillar palatal process and the palatine, and the alveolar processes of the maxilla can be recovered from the fossil material. On the other hand, the reconstruction of the braincase, ear region, and basicranial architecture is a matter of discussion. From the phylogenetic point of view, the skulls of evolved anagalids (*Anagalopsis* and *Anagale*, see Bohlin 1951; McKenna 1963), Eocene eurymylids (*Rhombomylus*, see Meng et al. 2003) or large mimotonids (*Gomphos*, see Asher et al. 2005) are the closest proxies for the whole skull restoration of the earliest Glires (Fig. 5.5).

The skull of earliest Glires bears already important morphological characteristics of this group. The muzzle is strong, relatively high, with large premaxillae, and well-developed maxillae hosting powerful incisors. The incisive foramen is elongated and relatively large. The anterior roots of the zygomatic arch in basal Glires (and even the rodentiaform *Tribosphenomys*) are not yet as well developed as in more advanced groups (e.g., *Rhombomylus*; see Meng et al. 2003, or modern Glires), and still slightly more posteriorly placed, although they are relatively anteriorly placed compared to other placental groups, which indicates the formation of the typical Glires masseter system.



**Fig. 5.5** Reconstruction of the cranial structure of *Heomys orientalis* from the middle Paleocene of Qianshan (Anhui Province, China). **a** The holotype specimen (IVPP V4321), anterior part of the skull with complete dentition, a pair of DI2 and P3–M3 (the sediment removed virtually, compare to Fig. 5.3b); **b** the mandible body (IVPP V4322), with p4–m3 and roots of p3; **c** reconstruction of the skull morphology based on the preserved cranial material and the skull architecture known in the closest fossil relatives. Drawings by Agnieszka Kapuścińska

When compared, the skulls of *Heomys* and *Mimotona* immediately display two different architectures characteristic of Glires, which were already present in the early/middle Paleocene. The “mimotonid type” has a more delicate, lower muzzle, a wider dental part with shorter hard palate, further adopted by Lagomorpha while the “eurymylid type” is characterized by a stronger and higher muzzle and a long hard palate, as seen in rodents.

The muzzle part of the ischyromyid skull resembles in general the skull of *Heomys*, although the whole ischyromyid skull is overall strongly elongated (see Anderson 2008).

Even the earliest fossil record of the early and middle Paleocene already shows a disparity in basal Glires which corresponds to further duplicidentate or lagomorph versus simplicidentate/rodent morphotype diversification. Whether this Paleocene diversity corresponds to true phylogenetic affiliations or is a manifestation of frequent in Glires mosaic evolution is a matter for further study. Most probably *Mimotona* and Lagomorpha affiliation should hold, but most of the eurymylid lineages are probably the evolutionary dead ends. There are at least three (or possibly four) different morphological groups within the Paleocene eurymylids, but their relationships to Rodentiaformes and further rodents of modern aspect are unclear. Most probably, the eurymylid radiation as a whole is an incipient manifestation of the Glires ability to produce quick microevolutionary changes, exemplified by rodents. On the other hand, the striking similarity between *Mimotona* and modern lagomorphs emphasizes the morphological conservatism of Duplicidentata, which may have its roots deeper at the basal Euarchontoglires.

**Acknowledgements** I thank Pierre Pontarotti for the invitation to participate in this contribution and Marie-Hélène Rome for her great patience throughout. I am indebted to Li Chuan-Kui, Li Qian, Ni Xijun, and Wang Yuan-Qing for access to the specimens and stimulating discussions on the

Glires. Thanks are extended to Agnieszka Kapuścińska for the drawings of the *Heomys* skull and to Aleksandra Hołda-Michalska for help with other figures. I acknowledge the support of the National Science Center (Poland) (grant No. 2015/18/E/NZ8/00637).

## References

- Anderson D (2008) Ischyromyidae. In: Janis CM, Gunnell GF, Uhen MD (eds) Evolution of tertiary mammals in North America. Small mammals, xenarthrans, and marine mammals, vol 2. Cambridge University Press, Cambridge, pp 311–325
- Asher RJ, Meng J, Wible JR, McKenna MC, Rougier GW, Dashzeveg D, Novacek MJ (2005) Stem lagomorpha and the antiquity of Glires. *Science* 307:1091–1094
- Asher RJ, Bennett N, Lehmann T (2009) The new framework for understanding placental mammal evolution. *BioEssays* 31:853–864
- Asher RJ, Smith MR, Rankin A, Emry RJ (2019) Congruence, fossils and the evolutionary tree of rodents and lagomorphs. *R Soc Open Sci* 6:190387
- Averianov AO (1994) Early Eocene mimotonids of Kyrgyzstan and the problem of Mixodontia. *Acta Palaeontol Pol* 39:393–411
- Bohlin B (1951) Some mammalian remains from Shih-ehr-ma-ch'eng, Hui-hui-p'u area, Western Kansu. Reports from the scientific expedition to the North-Western Provinces of China under leadership of Dr Sven Hedin. The Sino-Swedish Expedition Publication 35, VI. *Vert Pal* 5:1–48
- Chow MC, Qi T (1978) Paleocene mammalian fossils from Nomogen formation of Inner Mongolia. *Vert Pal Asiat* 16:77–85
- Dashzeveg D, Russell DE (1988) Palaeocene and Eocene Mixodontia (Mammalia, Glires) of Mongolia and China. *Palaeontology* 31:129–164
- Dashzeveg D, Hartenberger J, Martin T, Legendre S (1998) A peculiar minute Glires (Mammalia) from the early Eocene of Mongolia. *Bull Carnegie Mus Nat Hist* 34:194–209
- Dawson MR, Beard KC (1996) New late Paleocene rodents (Mammalia) from big multi quarry, Washakie Basin, Wyoming. *Palaeovertebrata* 25:301–321
- Flynn LJ (2008) Hystricognathi and Rodentia incertae sedis. In: Janis CM, Gunnell GF, Uhen MD (eds) Evolution of tertiary mammals in North America. Small mammals, xenarthrans, and marine mammals (vol 2). Cambridge University Press, Cambridge, pp 498–506
- Fostowicz-Frelik Ł (2016) A new zalambdalestid (Eutheria) from the Late Cretaceous of Mongolia and its implications for the origin of Glires. *Palaeontol Pol* 67:127–136
- Fostowicz-Frelik Ł (2017) Convergent and parallel evolution in early Glires (Mammalia). In: Pontarotti P (ed) *Evolutionary biology: self/nonself evolution, species and complex traits evolution, methods and concepts*. Springer, Cham, pp 199–216
- Fostowicz-Frelik Ł, Li CK, Mao FY, Meng J, Wang YQ (2015) A large mimotonid from the Middle Eocene of China sheds light on the evolution of lagomorphs and their kin. *Sci Rep* 5:9394
- Fostowicz-Frelik Ł, Li Q, Ni X (2018) Oldest ctenodactyloid tarsals from the Eocene of China and evolution of locomotor adaptations in early rodents. *BMC Evol Biol* 18:e150
- Huang XS, Li CK, Dawson MR, Liu LP (2004) *Hanomys malcolmi*, a new simplicidentate mammal from the Paleocene of central China: its relationships and stratigraphic implications. *Bull Carnegie Mus Nat Hist* 36:81–89
- Ivy LD (1990) Systematics of the late Paleocene and early Eocene Rodentia (Mammalia) from the Clarks Fork Basin, Wyoming. *Contrib Mus Pal U Mich* 28:21–70
- Kielan-Jaworowska Z, Cifelli RL, Luo ZX (2004) *Mammals from the age of dinosaurs: origins, evolution, and structure*. Columbia University Press, New York
- Li CK (1977) Paleocene eurymyloids (Anagalida, Mammalia) of Qianshan, Anhui. *Vert Pal Asiat* 15:103–118

- Li CK, Ting SY (1985) Possible phylogenetic relationship of Asiatic eurymylids and rodents, with comments on mimotonids. In: Lockett WP, Hartenberger JL (eds) Evolutionary relationships among rodents. Plenum Press, New York, pp 35–58
- Li CK, Ting SY (1993) New cranial and postcranial evidence for the affinities of the eurymylids (Rodentia) and mimotonids (Lagomorpha). In: Szalay FS, Novacek MJ, McKenna MC (eds) Mammal phylogeny—placentals. Springer, Berlin, pp 151–158
- Li CK, Meng J, Wang YQ (2007) *Dawsonolagus antiquus*, a primitive lagomorph from the Eocene Arshanto formation, Nei Mongol, China. Bull Carnegie Mus 39:97–110
- Li CK, Wang YQ, Zhang ZQ, Mao FY, Meng J (2016) A new mimotonidan mammal *Mina hui* (Mammalia, Glires) from the Middle Paleocene of Qianshan, Anhui Province, China. Vert Pal Asiat 54:121–136
- Lopatin AV, Averianov AO (2004a) The earliest rodents of the genus *Tribosphenomys* from the Paleocene of Central Asia. Dokl Biol Sci 397:336–337
- Lopatin AV, Averianov AO (2004b) A new species of *Tribosphenomys* (Mammalia: Rodentiaformes) from the Paleocene of Mongolia. New Mexico Mus Nat Hist Sci Bull 26:169–175
- Mao FY, Li Q, Wang YQ, Li CK (2017) *Taizimylus tongi*, a new eurymylid (Mammalia, Glires) from the upper Paleocene of Xinjiang, China. Palaeoworld 26:519–530
- Matthew WD, Granger W (1925) Fauna and correlation of the Gashato formation of Mongolia. Am Mus Novit 186:1–12
- McKenna MC (1963) New evidence against tupaioid affinities of the mammalian family Anagalidae. Am Mus Novit 2158:1–16
- McKenna MC, Meng J (2001) A primitive relative of rodents from the Chinese Paleocene. J Vert Pal 21:565–572
- Meng J, Wyss AR (2001) The morphology of *Tribosphenomys* (Rodentiaformes, Mammalia): phylogenetic implications for basal Glires. J Mamm Evol 8:1–71
- Meng J, Wyss AR (2005) Glires (Lagomorpha, Rodentia). In: Rose KD, Archibald JD (eds) The rise of placental mammals: origins and relationships of the major extant clades. Johns Hopkins University Press, Baltimore, pp 145–158
- Meng J, Wyss AR, Dawson MR, Zhai R (1994) Primitive fossil rodent from Inner Mongolia and its implications for mammalian phylogeny. Nature 370:134–136
- Meng J, Zhai RJ, Wyss AR (1998) The late Paleocene Bayan Ulan fauna of Inner Mongolia, China. Bull Carnegie Mus Nat Hist 34:148–185
- Meng J, Hu YM, Li CK (2003) The osteology of *Rhombomylus* (Mammalia, Glires): implications for phylogeny and evolution of Glires. Bull Am Mus Nat Hist 275:1–247
- Meng J, Wyss AR, Hu Y, Wang YQ, Bowen GJ, Koch PL (2005) Glires (Mammalia) from the late Paleocene Bayan Ulan locality of Inner Mongolia. Am Mus Novit 3473:1–25
- Meng J, Ni X, Li CK, Beard KC, Gebo DL, Wang YQ, Wang HJ (2007) New material of Alagomyidae (Mammalia, Glires) from the late Paleocene Subeng locality, Inner Mongolia. Am Mus Novit 3597:1–29
- Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351
- Rinderknecht A, Blanco RE (2008) The largest fossil rodent. Roy Soc Proc B 275:923–928
- Rose KD (1981) The Clarkforkian land-mammal age and mammalian composition across the Paleocene-Eocene boundary. Univ Michigan Pap Paleontol 26:1–197
- Rose KD, Chinnery BJ (2004) The postcranial skeleton of early rodents. Bull Carnegie Mus Nat Hist 36:211–244
- Smith AG, Smith DG, Funnell BM (1994) Atlas of Mesozoic and Cenozoic coastlines. Cambridge University Press, Cambridge
- Sych L (1971) Mixodontia, a new order of mammals from the Paleocene of Mongolia. Palaeontol Pol 25:147–158
- Voss RS (1988) Systematics and ecology of Ichthyomyine rodents (Muroides) patterns of morphological evolution in a small adaptive radiation. Bull Am Mus Nat Hist 188:259–493



- Wang YQ, Li CK, Li DS (2016) A synopsis of Paleocene stratigraphy and vertebrate paleontology in the Qianshan Basin, Anhui, China. *Vert Pal Asiat* 54:89–120
- Wilson DE, Lacher TE, Mittermeier RA (eds) (2016) *Handbook of the mammals of the World: lagomorphs and rodents I*. Barcelona Lynx Edicions
- Wood AE (1962) The early tertiary rodents of the family Paramyidae. *Trans Am Philos Soc* 52(1):1–261
- Wyss AR, Meng J (1996) Application of phylogenetic taxonomy to poorly resolved crown clades: a stem-modified node-based definition of Rodentia. *Syst Biol* 45:559–568

# Chapter 6

## Continuous Spectrum of Lifestyles of Plant-Associated Fungi Under Fluctuating Environments: What Genetic Components Determine the Lifestyle Transition?



Kei Hiruma

**Abstract** Plants interact with diverse fungal species, ranging from pathogens to beneficial endophytes. The pathogenic and beneficial lifestyles of fungi have often been studied separately and independently, so the aspects of genetic basis that contribute to lifestyle transitions in plant-associated fungi have not been generally addressed. The *Colletotrichum* genus comprises a highly diverse group of pathogens that infect and cause anthracnose diseases in a wide range of plant hosts. On the other hand, some of the *Colletotrichum* species act as beneficial endophytes and promote plant growth under conditions of stress. The presence of diverse *Colletotrichum* species with contrasting infection strategies thus provides a suitable model system in which to explore the molecular basis for discriminating pathogenic and beneficial lifestyles of plant-associated fungi. This chapter reviews recent molecular-based research related to pathogenic and beneficial *Colletotrichum* species and discusses the possible molecular basis underlying the lifestyle determination, based on the results of comparative genomics and *in planta* transcriptome analysis.

### 6.1 Introduction

Plants associate intimately with diverse microbes, ranging from pathogens causing disease to beneficial microbes promoting plant growth. Unlike animal guts, in which bacterial species are dominant, plants also host diverse eukaryotic fungal species. However, despite their richness and diversity in plant ecosystems, much less is known about the eco-physiological functions of fungal species than is understood for bacterial ones (Rodriguez et al. 2009). Nevertheless, several host and fungal genetic factors

---

K. Hiruma (✉)

Department of Science and Technology, Nara Institute of Science and Technology, Nara 630-0192, Japan

e-mail: [hiruma@bs.naist.jp](mailto:hiruma@bs.naist.jp)

PRESTO, Japan Science and Technology Agency, 4-1-8 Honcho Kawaguchi, Saitama 332-0012, Japan

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_6](https://doi.org/10.1007/978-3-030-57246-4_6)

117

required for pathogenic lifestyles of plant-associated fungi have been identified in several plant–pathogen interaction model systems (Boller and He 2009). Similarly, genetic factors underlying the lifestyles of beneficial fungi have been identified in the context of plant interactions with mutualistic arbuscular mycorrhizal fungi that promote plant growth under nutrient-limiting conditions or for some of the root-associated endophytes such as beneficial *Serendipita indica* (Bonfante and Genre 2010; Varma et al. 1999). However, as most of the molecular-level reports related to plant–microbe interactions have focused on specific details involved in each type of association, there has been little generalization about molecular mechanisms that are critical for a selection of lifestyle as either pathogens or mutualists.

Although fungal pathogenic and beneficial lifestyles appear to be quite different, it has been reported that closely related fungal species often behave with opposite lifestyles in the same host (Hacquard et al. 2016; de Lamo and Takken 2020), suggesting that subtle genetic differences determine the lifestyles of plant-associated fungi. Furthermore, some host factors have contributed to colonization by both pathogenic and beneficial fungi (Wang et al. 2012), suggesting the presence of a common pathway for plant-associated fungi. This is also consistent with the ecological view that lifestyles of plant-associated microbes sometimes show continuity from pathogens to mutualists, depending on the host and environmental conditions (Hardoim et al. 2015). Understanding the basis for how lifestyles of microbes are determined in hosts will break down of the dogma of pathogen or mutualistic lifestyles labels, into a more continuous spectrum of lifestyle interactions.

The ascomycete genus *Colletotrichum* causes anthracnose diseases in a wide range of economically important crops, and is considered to be one of the top 10 most devastating fungal pathogens of scientific and economic importance (Dean et al. 2012). Interestingly, some of the *Colletotrichum* species are reported as saprotrophs and also as endophytes that colonize plant tissues without causing disease symptoms. The whole genome information and/or the in-depth *in planta* transcriptome data for pathogenic and endophytic *Colletotrichum* fungi have been reported (O’Connell et al. 2012; Gan et al. 2013; Hiruma et al. 2016). Thus, the accumulated information on the diverse lifestyles of *Colletotrichum* species can help to elucidate the genetic basis discriminating pathogenic and endophytic lifestyles of plant-associated fungi. This chapter first briefly summarizes the published reports on molecular examination of pathogenic and endophytic *Colletotrichum* species. Based on available comparative genome and *in planta* transcriptome analysis, the chapter then considers a possible genetic basis that can discriminate pathogenic and endophytic lifestyles of *Colletotrichum*, as well as future perspectives for identification of a genetic basis for these tendencies.

## 6.2 *Colletotrichum* Fungal Species as Pathogens

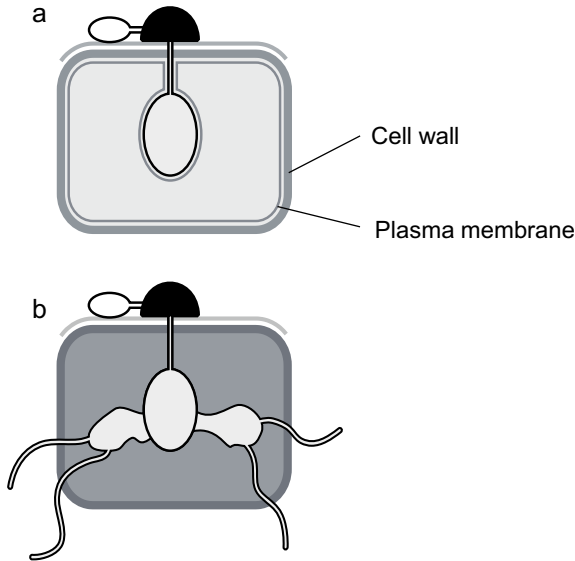
The large ascomycete genus *Colletotrichum* causes anthracnose diseases in a wide range of economically important crops, and has been named among the top 10 devastating fungal pathogens of scientific and economic importance (Dean et al. 2012). Many *Colletotrichum* species, as well as the rice blast fungus *Magnaporthe oryzae*, undertake a hemibiotrophic infection strategy after their invasion into the host tissues, in which an initial biotrophic phase dependent on living host cells is followed by a destructive necrotrophic phase (Perfect et al. 1999). In contrast to genuine obligate biotrophs such as powdery mildew and arbuscular mycorrhizal fungi, most of the already-described *Colletotrichum* species are readily amenable to axenic culture and genetic manipulation, which enables access to functional fungal genetic analysis. Added to this, high-quality genome sequences are available for more than 10 *Colletotrichum* species, which facilitates comparative genomics and molecular genetic studies in this fungal genus (Table 6.1, O'Connell et al. 2012; Gan et al. 2013, 2016; Hacquard et al. 2016).

Soon after a spore attaches to the surface of its host, the spore starts its morphological developmental process, along with secretion of various enzymes for host invasion (Tucker and Talbot 2001). For fungal entry into host leaves, the majority of reported *Colletotrichum* and other hemibiotrophic fungal pathogens such as the rice blast fungus *Magnaporthe oryzae* species form dome-shaped infection structures called appressoria soon after a spore attaches to the surface of its host (Ryder and Talbot 2015). The infection structures are heavily black-melanized, and are considered to enable the pathogenic fungus to generate enough turgor pressure (up to 6–8 MPa) to invade host tissues protected by tight cell-wall components (Kubo and Furusawa 1991; Howard and Valent 1996; Dean 1997; de Jong et al. 1997). Indeed, genetic manipulation or chemicals targeted at the melanin biosynthetic pathway are effective in inhibiting such entry and pathogenesis (Hiruma et al. 2010). Plant infection by pathogens involves secretion of effector proteins that suppress plant immunity responses and facilitate pathogen growth within plant tissues (O'Connell et al. 2012; Lo Presti et al. 2015). Interestingly, development of appressoria on host tissues is also tightly linked with secretion of effectors. It has been reported that virulence-related effectors of *Colletotrichum higginsianum* and *C. orbiculare* are focally accumulated at appressorial penetration pores (Kleemann et al. 2012), suggesting that the effectors are secreted from the pores. Infection-stage specific transcriptome analysis during leaf colonization by pathogenic *C. higginsianum* also revealed that genes encoding cell-wall-degrading enzymes are also upregulated (O'Connell et al. 2012), suggesting that pathogenic *Colletotrichum* penetrates the thick plant cell wall through the use of enzymes that degrade the host plant cell walls.

After penetration, the intracellular hyphae of most of the characterized *Colletotrichum* fungi are enclosed by the host membrane and establish a transient biotrophic phase with the host plant (Fig. 6.1). Analysis of the transcriptome during leaf colonization by pathogenic *C. higginsianum* revealed that several genes related

**Table 6.1** Lists of representative available *Colletotrichum* whole genome information

Species	Strains	Life-styles	Hosts	References	Clade
<i>C. fiorinia</i>	PJ7	Pathogen	Varioius plants	Baroncelli et al. (2014a, b)	Acutatum
<i>C. orchidophilum</i>	IMI 309357	Pathogen	Orchid	Baroncelli et al. (2018)	Acutatum
<i>C. salicis</i>	CBS 607.94	Pathogen	Varioius plants	Baroncelli et al. (2016)	Acutatum
<i>C. simmondsii</i>	CBS122122	Pathogen	Varioius plants	Baroncelli et al. (2016)	Acutatum
<i>C. acutatum</i>	KC05	Pathogen	Peper	Han et al. (2016)	Acutatum
<i>C. graminicola</i>	M1.001	Pathogen	Maize	O'Connell et al. (2012)	Graminicola
<i>C. sublineola</i>	TX430BB	Pathogen	Sorghum	Baroncelli et al. (2014)	Graminicola
<i>C. incanum</i>	MAFF 238704, MAFF 238706, MAFF238712, MAFF238713	Pathogen	Radish, <i>A. thaliana</i> , lily	Gan et al. (2017), Hacquard et al. (2016)	Spaethianum
<i>C. tofieldiae</i>	0861, CBS168.49, CBS130851, CBS 495.85	Endophyte	<i>A. thaliana</i>	Hacquard et al. (2016)	Spaethianum
<i>C. higginsianum</i>	IMI 349063, MAFF 305635	Pathogen	<i>A. thaliana</i>	O'Connell et al. (2012), Dallery et al. (2017), Tsushima et al. (2019)	Destructivum
<i>C. tanacetii</i>	BRIP57314	Pathogen	Pyrethrum	Lelwala et al. 2019	Destructivum
<i>C. shisoi</i>		Pathogen	<i>Perilla frutescens</i>	Gan et al. 2019	Destructivum
<i>C. chlorophyti</i>	NTL11	Pathogen	Legumes, tomato, soybean	Gan et al. (2017)	
<i>C. fruticola</i>	Nara-gc5	Pathogen	Strawerry	Gan et al. (2013)	Gloeosporioides
<i>C. fruticola</i>	CGMCC3.17371	Pathogen	Strawerry	Armitage et al. (2020)	Gloeosporioides
<i>C. fruticola</i>	1104-7	Pathogen	Apple	Liang et al. (2018)	Gloeosporioides
<i>C. gloeosporioides</i>	Cg-14	Pathogen	Vairous Fruits	Alkan et al. 2013	Gloeosporioides
<i>C. truncatum</i>	MTCC no. 3414	Pathogen	chilli	Rao and Nandineni (2017)	Truncatum
<i>C. orbiculare</i>	104-T	Pathogen	Cucumber	Gan et al. (2013)	Orbiculare
<i>C. lindemuthianum</i>		Pathogen	Bean	de Queiroz et al. (2017)	Orbiculare



**Fig. 6.1** Leaf infection process by pathogenic *Colletotrichum* species. **a** Majority of pathogenic *Colletotrichum* species form dome-shaped black-melanized appressoria on leaf surface soon after the spores land in the surface. Via turgor pressure and cell-wall-degrading enzymes, the pathogens penetrate host cells and form biotrophic hyphae that are enclosed by host plasma membrane in epidermal cells. Yellow color represents cuticle layer. **b** After transient biotrophic phase, pathogenic *Colletotrichum* species turns to a necrotrophic phase during which the pathogens develop the thinner hyphae (than biotrophic hyphae) and actively kill host cells. The transition timing from biotrophic phase to necrotrophic phase is diversified among *Colletotrichum*

to effectors, which are different from genes induced during penetration, are specifically induced during biotrophic interactions. Maximum numbers of the effector candidate genes are highly induced during the biotrophic phase, so the biotrophic interface (between the fungal hyphae and plant membrane component) appears to be a site for such effector secretion. In support of this idea, virulence-related effectors of pathogenic *C. orbiculare* fused with fluorescence protein accumulated in a ring-like region around the neck of the primary biotrophic hyphae in a manner dependent on an exocytosis-related component, namely, Rab GTPase SEC4 (Irieda et al. 2014). Combining this observation with the fact that disruption of SEC4 attenuates the virulence of *C. orbiculare*, it appears that virulence effectors are secreted via ring-like regions at the interface. Some virulence-related effectors of hemibiotrophic *Magnaporthe oryzae* focally accumulate in the biotrophic interfacial complex formed in a space between the plant membrane and biotrophic hyphae, which are different from the ring-like regions formed in biotrophic hyphae of *C. orbiculare* (Giraldo et al. 2013). These findings suggest that the mechanisms of effector delivery via fungal biotrophic hyphae could be diverse among hemibiotrophic pathogens. This contrasts with the case of well-conserved pathogenic bacterial strategies to inject effectors into the cytosol of eukaryotic cells via a type III secretion system (Hueck 1998).

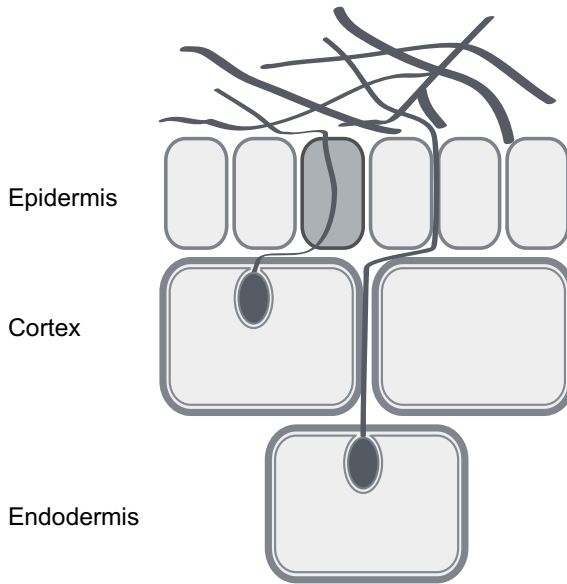
It would be interesting to investigate whether formation of ring-like structures is specific to *C. orbiculare* or is widespread across the *Colletotrichum* genus. In addition to effectors, it is noteworthy that several genes related to secondary metabolism are also highly up-regulated during the biotrophy (O'Connell et al. 2012, See also the section of "Repertoires of secondary metabolites in pathogenic versus beneficial *Colletotrichum*").

The biotrophic phase is transient and suddenly switches to a neurotrophic phase. During the neurotrophic phase, pathogenic *Colletotrichum* fungi differentiate thin, rapidly growing hyphae that kill host tissues (Fig. 6.1). During the necrotrophic phase, *Colletotrichum* pathogens induce genes encoding cell-wall-degrading enzymes that target different types of plant cell-wall components: proteases, necrosis-inducing proteins, and secondary metabolites including several putative fungal toxins (O'Connell et al. 2012). This transcriptomic reprogramming during necrotrophy appears to be adapted to kill host cells and to retrieve nutrients from host tissues. However, the mechanisms by which fungi start transition from late biotrophy to necrotrophy are not clear.

### 6.3 *Colletotrichum* Species as Endophytes

In addition to hemibiotrophic pathogenic species, the *Colletotrichum* genus has endophytic *Colletotrichum* species that colonize inside plant tissues without causing disease symptoms and that in some cases confer benefits to the host plants. Many reports show that various *Colletotrichum* species have been isolated from various healthy plants after surface disinfection, suggesting these strains are endophytes living inside host tissues without causing visible disease, at least as applies to the moment when they were isolated. However, infection processes of most of the isolated putative endophytic *Colletotrichum* species have not been further characterized in the laboratory or in the field, so it is still not clear whether the *Colletotrichum* fungi isolated from healthy plants after surface disinfection associate with host plants as true endophytes or just as stochastic encounters or as pathogens for which the virulence is suppressed via mechanisms as yet unknown. Importantly, however, it has been reported that *C. tofieldiae* isolated from and frequently detected in several different healthy wild *Arabidopsis thaliana* populations in Spain asymptotically colonizes the host roots and, importantly, promotes plant growth under low-phosphate conditions, in part by transferring phosphorus to the host via the hyphae (Fig. 6.2, Hiruma et al. 2016). This resembles the action of arbuscular mycorrhizal fungi that promote plant growth via nutrient transfer (Bonfante and Genre 2010). In contrast to arbuscular mycorrhizal fungi that receive carbon from host plants, however, what kinds of benefits *C. tofieldiae* receive when the fungus transfers phosphorus to the host plants is still an open question.

Interestingly, in-depth microscopic analysis revealed that although *C. tofieldiae* causes epidermal cell death after the transient biotrophic phase, *C. tofieldiae* appears to form a stable biotrophic interface with host plants in cortex cells, the second



**Fig. 6.2** Root infection process by the beneficial endophytic *Colletotrichum tofieldiae* (*Ct*). Hyphae of *Ct* start to penetrate epidermal root cells without forming appressoria. After transient biotrophic interactions in epidermal cells, the infected epidermal cells lost viability. Hyphae of *Ct* also localizes in intercellular regions. In cortex or endodermis, hyphae of *Ct* are enclosed by host plasma membrane and appears to form stable biotrophic interactions. The transition from biotrophy to necrotrophy has not been observed in *A. thaliana* Col-0 plants

layers of plant root cells. The morphological features of the interface are similar to those of the transient biotrophic interface formed by the majority of pathogenic *Colletotrichum*. However, transition from biotrophy to necrotrophy, as observed in most of the pathogenic *Colletotrichum*, is not observed during *C. tofieldiae* root colonization, and the absence of this transition might thus prevent disease symptoms during fungal colonization. At the genome level, however, *C. tofieldiae* is very closely related to root-infecting pathogenic species belonging to the spathianum clade, such as *C. incanum* (Hacquard et al. 2016). Indeed, even *C. tofieldiae* displays high virulence in host *cyp79B2 cyp79B3* mutant plants lacking host tryptophan-derived antimicrobial metabolites, including the phytoalexin camalexin and indole glucosinolates (Hiruma 2019; Hiruma et al. 2016). This in turn suggests that the antifungal metabolite pathway is required to suppress the potential pathogenesis of *C. tofieldiae*.

There are diverse ranges of infection strategies that range from pathogenic to beneficial in *Colletotrichum* species. What is the molecular basis that discriminates pathogenic from beneficial infection strategies? The infection strategies for pathogenic and beneficial action appear to be quite different. However, it has been reported that deletion of one genetic locus turns pathogenic *C. magna* to an endophyte that protects the host plants from pathogens (Freeman and Rodriguez 1993; Redman



et al. 1999). Similarly, deletion of one plant tryptophan-derived metabolite pathway is enough to turn beneficial *C. tofieldiae* or another beneficial endophyte *Serendipita indica* to a pathogen, under the condition where the beneficial fungi promotes the growth of some wild-type plants (Nongbri et al. 2012; Lahrmann et al. 2015; Hiruma et al. 2016). These results suggest that lifestyles of plant-associated *Colletotrichum* could be determined using a tractable genetic basis. Such high phylogenetic relatedness to pathogenic species seems to be common, rather than exceptional, for endophytic fungi isolated from healthy, surface-sterilized tissues of different plant species (Rodriguez et al. 2009). Thus, as a first step to identifying the genetic basis discriminating beneficial and pathogenic lifestyles, it is very useful to perform comparative analysis that includes cytological, genomic, and *in planta* transcriptome analysis using *Colletotrichum* species with diverse lifestyles as a model (Table 6.1).

#### 6.4 Apparent Lack of a Transition from Biotrophy to Necrotrophy in Beneficial *Colletotrichum*

As described in the preceding sections, microscopic analysis suggests that beneficial *C. tofieldiae* does not show a transition from biotrophy to necrotrophy, which contrasts with a few characterized root-infecting pathogenic *Colletotrichum* species (Sukno et al. 2008). Arbuscular mycorrhizal fungi also form a biotrophic interface called arbuscule in cortex cells and do not show transition to any necrotrophy (Bonfante and Genre 2010). Rather, the old arbuscule is degraded by host plants (Kobae et al. 2016). Although the molecular mechanisms underlying the transition from biotrophy to necrotrophy are not yet understood, *C. higginsianum* mutants defective in pathogenicity as a result of *Agrobacterium*-mediated insertion of T-DNA in a genomic region do not show a transition from biotrophy to necrotrophy (Huser et al. 2009). One *C. higginsianum* mutant lacking a mini chromosome also failed to switch to necrotrophy (Plaumann et al. 2018). It is not currently clear what defects cause the *C. higginsianum* mutants to fail to switch to necrotrophy. Interestingly, a recent study suggests that penetration hyphae of *C. orbiculare* mutants lacking the homeobox transcription factor CoHox1 did not turn to necrotrophy even after 19 days post inoculation (Yokoyama et al. 2019), suggesting that CoHox1 is involved in the transition from biotrophy to necrotrophy. It is also interesting to address whether the prolonged biotrophic phase observed in the mutants causes any effects on plant growth and development, especially under stress conditions in which beneficial fungi often provide fitness benefits to host plants. A more detailed analysis of these nonpathogenic mutants, including time-resolved RNAseq analysis, as well as discovering the causative genes responsible for the phenotype, will also help investigators to understand how beneficial fungi regulate the transition (restrict necrotrophy), despite the fact that they share very similar genomes with relatively pathogenic *Colletotrichum*.

## 6.5 Repertoires of Cell-Wall-Degrading Enzymes in Pathogenic Versus Beneficial *Colletotrichum*

The plant cell wall, composed of a matrix of pectin, hemicellulose, lignin, and structural proteins, is a barrier that prevents pathogen infection (Kubicek et al. 2014). Hemibiotrophic pathogens induce sets of cell-wall-degrading enzymes during host infection (O'Connell et al. 2012). Comparative genomics between pathogenic *Colletotrichum* fungi show that repertoires of cell-wall-degrading enzymes are different, depending on which host plants the pathogens preferentially colonize (King et al. 2011; O'Connell et al. 2012). For example, *C. higginsianum* encodes more than twice as many pectin-degrading enzymes as does *C. gramminicola*, which appears to be well reflected in their host preferences for plants (dicot plants have more pectin than monocot plants). In contrast, most of the characterized beneficial fungi such as arbuscular mycorrhizal fungi and ectomycorrhizal fungi have reduced repertoires of genes encoding cell-wall-degrading enzymes (Tisserant et al. 2013; Nagendran et al. 2009; Martin et al. 2008). However, comparative genomic analysis of beneficial *C. tofieldiae* versus pathogenic *C. incanum*, both of which infect roots of *A. thaliana*, reveals that beneficial *C. tofieldiae* have similar repertoires of cell-wall-degrading enzymes. Restriction of repertoires of cell-wall-degrading enzymes has not been described for other beneficial endophytic fungi such as *Serendipita indica*, *Harpophora oryzae*, and *Helotiales* (Zuccaro et al. 2011; Xu et al. 2014; Almario et al. 2017). Furthermore, *in planta* transcriptome analysis has revealed that beneficial *C. tofieldiae* strongly expresses fungal genes encoding cell-wall-degrading enzymes during the root colonization, especially during the late colonization phase, during which *C. tofieldiae*-mediated plant growth promotion is clearly detected. These gene-encoded cell-wall-degrading enzymes act on all major polymers, including cellulose, hemicellulose, and pectin (Hacquard et al. 2016). Considering that *C. tofieldiae* promotes seed production as well, these results suggest that the absence or presence of these genes might be not key in distinguishing pathogenic from beneficial, at least in the case of the *Colletotrichum*. Unlike the case of obligate biotrophy as in arbuscular mycorrhizal fungi, the characterized endophytes including *C. tofieldiae* can be grown in the absence of host plants, and its colonization in host plants partially induced cell death, such as in epidermal cell layers (Deshmukh et al. 2006; Hiruma et al. 2016). Thus, the contrasting differences between biotrophs and hemibiotrophs might rather determine the differences in repertoires and *in planta* induction of cell-wall-degrading enzymes. The roles of cell-wall-degrading enzymes during root colonization by beneficial fungi merit future in-depth studies.

## 6.6 Repertories of Secondary Metabolites in Pathogenic Versus Beneficial *Colletotrichum*

Filamentous fungi produce diverse secondary metabolites. Fungal secondary metabolites are of intense interest due to their pharmaceutical (antibiotics) and/or toxic (mycotoxins) properties (Yu and Keller 2005). Especially, whole genome analysis of *Colletotrichum* has demonstrated the high potential of *Colletotrichum* species for secondary metabolite production compared with relative plant-associated fungi, a factor that can be assumed from the higher numbers of secondary metabolism clusters (O'Connell et al. 2012). More than 100 secondary metabolites have been isolated from pathogenic and endophytic *Colletotrichum* species (Kim and Shim 2019). For example, several antimicrobial compounds, and plant hormones such as indole-3-acetic acid, have been isolated from a liquid culture of several *Colletotrichum* species (Zou et al. 2000; Lu et al. 2000). In terms of the regulatory mechanisms of secondary metabolism clusters, it was recently reported that some of the secondary metabolism clusters have been silenced in the absence of host plants via H3K4 trimethylation (Dallery et al. 2019), a finding that appears similar to results in *Aspergilli* and *Saccharomyces cerevisiae* (Bok et al. 2009; Palmer et al. 2013; Shinohara et al. 2016). However, although fungal mutants can lose the ability to produce particular secondary metabolites and can lose the ability to infect, little is so far known about the functions of such secondary metabolites during infection. Interestingly, exogenous application of higginsianin B, produced from one of the secondary metabolism clusters regulated by H3K4 trimethylation in pathogenic *C. higginsianum* during the penetration to biotrophic phase, suppresses jasmonate-mediated plant defenses, likely via inhibition of 26S proteasome-dependent degradation of JAZ proteins (Dallery et al. 2020). Importantly, Markov cluster algorithm (MCL) analysis comparing *C. tofieldiae* and its pathogenic relative *C. incanum* revealed that the beneficial *C. tofieldiae* possesses many more gene families for secondary metabolite biosynthesis (Hacquard et al. 2016), implying critical roles for secondary metabolites in the lifestyles of beneficial fungi.

## 6.7 Repertories of Candidate Effectors, Comparing Pathogenic Versus Beneficial *Colletotrichum*

During the evolutionary arms race between host plants and pathogens, plant-associated pathogens have developed several different effectors to effectively suppress host defense responses (see Hogenhout et al. 2009; Raffaele and Kamoun 2012; Sanchez-Vallet et al. 2018 evolutionary development of pathogenic effectors from genome aspect). The mutualistic arbuscular mycorrhizal fungi and ectomycorrhizal fungi also use effectors to manipulate the host hormonal pathway to promote infection (Kloppholz et al. 2011; Plett et al. 2011). Thus, it is conceivable that plant-associated microbes may preferably increase the diversity of effector

repertoires to effectively suppress host defense responses. However, the numbers of annotated candidate effectors in beneficial *C. tofieldiae* are smaller than those in pathogenic *C. incanum* (Hacquard et al. 2016). In addition, time-resolved *in planta* transcriptome analysis has suggested that activation of effector genes in *C. tofieldiae* is weaker than for those of *C. incanum*. Reduced repertoires of candidate effectors in genomes of endophytes compared with those of pathogens have been reported also in *Fusarium* (de Lamo and Takken 2020). As introduced in the preceding sections, effectors have a necessary role in colonizing host tissues by suppressing plant immunity and other responses. At the same time, host plants have developed resistance genes encoding nucleotide-binding leucine-rich repeat proteins (NLRs) to directly or indirectly detect the activities of effectors for termination of pathogen growth in host tissues (Bergelson et al. 2001). Indeed, it has been reported that only two plant NLRs namely ZAR1 and CAR1 in *Arabidopsis thaliana* are predicted to be responsible for detection of the majority of bacterial type III secreted effectors (more than 90%) that are distributed in populations of pathogen *Pseudomonas syringae* (Laflamme et al. 2020). This proposes unexpected broad spectrum defense responses against pathogens via effector recognition by NLRs. In addition, some of the effectors expressed during biotrophic to necrotrophic stages cause cell death responses when they are overexpressed in fungi or in plants. For example, overexpression of a biotrophy-specific *C. truncatum* effector in *C. truncatum* or in the rice blast pathogen *Magnaporthe oryzae* causes incompatibility with the host lentil and barley plants, respectively, by inducing cell death responses in infected host cells with the biotrophic hyphae (Bhadauria et al. 2013). Overexpression of *NIS1* of *C. orbiculare*, which is expressed during the biotrophic phase and has roles in suppressing plant immunity, induces cell death responses in *Nicotiana benthamiana* (Yoshino et al. 2012; Irieda et al. 2019). Overexpression of necrosis-inducing factors of *C. higginsianum*, which is expressed during necrotrophic phase induces necrotic cell death responses in *Nicotiana benthamiana* (Kleemann et al. 2012). It is not clear whether the effector-mediated cell death responses are among the essential functions of the effectors or are results of counterdefense responses by host plants, possibly via NLRs. In any case, the fact that transitions from biotrophy to necrotrophy have not been observed during wild-type *Arabidopsis thaliana* root colonization by *C. tofieldiae* suggests that the beneficial *C. tofieldiae* reduces the repertoire and the expression of effectors that directly or indirectly cause cell death responses to prevent unnecessary cell death responses in host plants. However, the *C. tofieldiae* turns into a pathogen with apparent necrotrophic growth in *Arabidopsis thaliana* plants lacking tryptophan-derived secondary metabolism, including antifungal indole glucosinolates and Camalexin (Hiruma et al. 2016), suggesting that *C. tofieldiae* retains at least minimum sets of effector repertoires to cause disease symptoms in susceptible host plants. It is important to address how the beneficial *C. tofieldiae* colonizes host roots and promotes plant growth with smaller repertoires of effectors, as well as weaker expression of the effector genes, than the relative pathogen *C. incanum*. It is also important to investigate the mechanisms by which the majority of candidate effectors (more than 100) in the *C. tofieldiae* genome remain silent during host infection.

## 6.8 Conclusion

As introduced in the preceding sections, comparative genomics and *in planta* transcriptome analysis between pathogenic and beneficial *Colletotrichum* species hint at the molecular mechanisms that discriminate pathogenic and beneficial lifestyles of fungi. To validate whether these candidates are really involved in transition of lifestyles of fungi is an important task in the future. For functional analysis, establishing a mutualistic, pathogenic, or beneficial plant–microbe association model in *A. thaliana* would promote a detailed and precise understanding of the mechanisms. In bacteria, comparative genomics between several pathogenic, commensal, and beneficial *Pseudomonas* species in *Arabidopsis thaliana* have identified that a bacterial factor shared among *Pseudomonas* species, probably via horizontal transfer, is responsible for determination of the lifestyles (Melnyk et al. 2019). However, due to the fact that very limited numbers of beneficial fungal endophytes have so far been well characterized for in-depth comparative analysis between pathogenic and beneficial fungi (e.g., Table 6.1, in *Colletotrichum*), little is known about factors determining lifestyles of fungi. Therefore, there is an urgent need to identify and characterize more endophytic fungi in order to obtain insights from comparative genomics and *in planta* transcriptome analysis. Compared to animals, plants have unique features that allow them to intimately associate with eukaryotic fungi. Understanding the still largely hidden lifestyles of plant-associated fungal species will increase the understanding of plant stress adaptation strategies and of molecular mechanisms that determine the evolutionary origin of pathogenesis and mutualism.

**Acknowledgements** I thank Akemi Uchiyama for help with figure construction. This work was supported in part by the Japan Society for the Promotion of Sciences (JSPS) KAKENHI Grant (20H02986), and the Japan Science and Technology Agency grant (JPMJPR16Q7).

## References

- Alkan N et al (2013) Global aspects of *pacC* regulation of pathogenicity genes in *Colletotrichum gloeosporioides* as revealed by transcriptome analysis. *Mol Plant Microbe Interact* 26:1345–1358. <https://doi.org/10.1094/Mpmi-03-13-0080-R>
- Almario J et al (2017) Root-associated fungal microbiota of nonmycorrhizal *Arabidopsis thaliana* and its contribution to plant phosphorus nutrition. *Proc Natl Acad Sci U S A* 114:E9403–E9412. <https://doi.org/10.1073/pnas.1710455114>
- Armitage AD et al (2020) Draft genome sequence of the strawberry anthracnose pathogen *Colletotrichum fructicola*. *Microbiol Resour Announc* 9. <https://doi.org/10.1128/MRA.01598-19>
- Baroncelli R, Sanz-Martin JM, Rech GE, Sukno SA, Thon MR (2014a) Draft genome sequence of *Colletotrichum sublineola*, a destructive pathogen of cultivated sorghum. *Genome Announc* 2. <https://doi.org/10.1128/genomeA.00540-14>
- Baroncelli R, Sreenivasaprasad S, Sukno SA, Thon MR, Holub E (2014b) Draft genome sequence of *Colletotrichum acutatum* Sensu Lato (*Colletotrichum fioriniae*). *Genome Announc* 2. <https://doi.org/10.1128/genomeA.00112-14>

- Baroncelli R et al (2016) Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. *BMC Genomics* 17:555. <https://doi.org/10.1186/s12864-016-2917-6>
- Baroncelli R et al (2018) Whole-genome sequence of the orchid anthracnose pathogen *Colletotrichum orchidophilum*. *Mol Plant Microbe Interact* 31:979–981. <https://doi.org/10.1094/MPMI-03-18-0055-A>
- Bergelson J, Kreitman M, Stahl EA, Tian DC (2001) Evolutionary dynamics of plant R-genes. *Science* 292:2281–2285. <https://doi.org/10.1126/science.1061337>
- Bhadauria V, Banniza S, Vandenberg A, Selvaraj G, Wei YD (2013) Overexpression of a novel biotrophy-specific *Colletotrichum truncatum* effector, CtNUDIX, in hemibiotrophic fungal phytopathogens causes incompatibility with their host plants. *Eukaryot Cell* 12:2–11. <https://doi.org/10.1128/Ec.00192-12>
- Bok JW et al (2009) Chromatin-level regulation of biosynthetic gene clusters. *Nat Chem Biol* 5:462–464. <https://doi.org/10.1038/nchembio.177>
- Boller T, He SY (2009) Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* 324:742–744. <https://doi.org/10.1126/science.1171647>
- Bonfante P, Genre A (2010) Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. *Nat Commun* 1:48. <https://doi.org/10.1038/ncomms1046>
- Dallery JF et al (2017) Gapless genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. *BMC Genomics* 18:667. <https://doi.org/10.1186/s12864-017-4083-x>
- Dallery JF et al (2019) H3K4 trimethylation by CclA regulates pathogenicity and the production of three families of terpenoid secondary metabolites in *Colletotrichum higginsianum*. *Mol Plant Pathol* 20:831–842. <https://doi.org/10.1111/mpp.12795>
- Dallery JF et al (2020) Inhibition of jasmonate-mediated plant defences by the fungal metabolite higginsianin B. *J Exp Bot* 71:2910–2921. <https://doi.org/10.1093/jxb/eraa061>
- de Jong JC, McCormack BJ, Smirnov N, Talbot NJ (1997) Glycerol generates turgor in rice blast. *Nature* 389:244–245. <https://doi.org/10.1038/38418>
- de Lamo FJ, Takken FLW (2020) Biocontrol by *Fusarium oxysporum* using endophyte-mediated resistance. *Front Plant Sci* 11:37. <https://doi.org/10.3389/fpls.2020.00037>
- de Queiroz CB, Correia HLN, Menicucci RP, Vidigal PMP, de Queiroz MV (2017) Draft genome sequences of two isolates of *Colletotrichum lindemuthianum*, the causal agent of anthracnose in common beans. *Genome Announc* 5. <https://doi.org/10.1128/genomeA.00214-17>
- Dean RA (1997) Signal pathways and appressorium morphogenesis. *Annu Rev Phytopathol* 35:211–234. <https://doi.org/10.1146/annurev.phyto.35.1.211>
- Dean R et al (2012) The Top 10 fungal pathogens in molecular plant pathology. *Mol Plant Pathol* 13:414–430. <https://doi.org/10.1111/j.1364-3703.2011.00783.x>
- Deshmukh S et al (2006) The root endophytic fungus *Piriformospora indica* requires host cell death for proliferation during mutualistic symbiosis with barley. *Proc Natl Acad Sci U S A* 103:18450–18457. <https://doi.org/10.1073/pnas.0605697103>
- Freeman S, Rodriguez RJ (1993) Genetic conversion of a fungal plant pathogen to a nonpathogenic, endophytic mutualist. *Science* 260:75–78. <https://doi.org/10.1126/science.260.5104.75>
- Gan P et al (2016) Genus-wide comparative genome analyses of colletotrichum species reveal specific gene family losses and gains during adaptation to specific infection lifestyles. *Genome Biol Evol* 8:1467–1481. <https://doi.org/10.1093/gbe/evw089>
- Gan P et al (2017) Draft genome assembly of *Colletotrichum chlorophyti*, a pathogen of herbaceous plants. *Genome Announc* 5. <https://doi.org/10.1128/genomeA.01733-16>
- Gan P et al (2019) *Colletotrichum shioi* sp. nov., an anthracnose pathogen of *Perilla frutescens* in Japan: molecular phylogenetic, morphological and genomic evidence. *Sci Rep-UK* 9:13349. <https://doi.org/10.1038/s41598-019-50076-5>

- Gan P et al (2013) Comparative genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of *Colletotrichum* fungi. *New Phytol* 197:1236–1249. <https://doi.org/10.1111/nph.12085>
- Giraldo MC et al (2013) Two distinct secretion systems facilitate tissue invasion by the rice blast fungus *Magnaporthe oryzae*. *Nat Commun* 4:1996. <https://doi.org/10.1038/ncomms2996>
- Hacquard S et al (2016) Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nat Commun* 7:11362. <https://doi.org/10.1038/ncomms11362>
- Han JH et al (2016) Whole genome sequence and genome annotation of *Colletotrichum acutatum*, causal agent of anthracnose in pepper plants in South Korea. *Genom Data* 8:45–46. <https://doi.org/10.1016/j.gdata.2016.03.007>
- Hardoim PR et al (2015) The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. *Microbiol Mol Biol R* 79:293–320. <https://doi.org/10.1128/Mmbr.00050-14>
- Hiruma K (2019) Roles of plant-derived secondary metabolites during interactions with pathogenic and beneficial microbes under conditions of environmental stress. *Microorganisms* 7:362. <https://doi.org/10.3390/microorganisms7090362>
- Hiruma K et al (2010) Entry mode-dependent function of an indole glucosinolate pathway in *Arabidopsis* for nonhost resistance against anthracnose pathogens. *Plant Cell* 22:2429–2443. <https://doi.org/10.1105/tpc.110.074344>
- Hiruma K et al (2016) Root endophyte *Colletotrichum tofieldiae* confers plant fitness benefits that are phosphate status dependent. *Cell* 165:464–474. <https://doi.org/10.1016/j.cell.2016.02.028>
- Hogenhout SA, Van der Hoorn RAL, Terauchi R, Kamoun S (2009) Emerging concepts in effector biology of plant-associated organisms. *Mol Plant Microbe Interact* 22:115–122. <https://doi.org/10.1094/Mpmi-22-2-0115>
- Howard RJ, Valent B (1996) Breaking and entering: host penetration by the fungal rice blast pathogen *Magnaporthe grisea*. *Annu Rev Microbiol* 50:491–512. <https://doi.org/10.1146/annurev.micro.50.1.491>
- Hueck CJ (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol R* 62:379–433. <https://doi.org/10.1128/Mmbr.62.2.379-433.1998>
- Huser A, Takahara H, Schmalenbach W, O'Connell R (2009) Discovery of pathogenicity genes in the crucifer anthracnose fungus *Colletotrichum higginsianum*, using random insertional mutagenesis. *Mol Plant Microbe Interact* 22:143–156. <https://doi.org/10.1094/Mpmi-22-2-0143>
- Irieda H et al (2014) *Colletotrichum orbiculare* secretes virulence effectors to a biotrophic interface at the primary hyphal neck via exocytosis coupled with SEC22-mediated traffic. *Plant Cell* 26:2265–2281. <https://doi.org/10.1105/tpc.113.120600>
- Irieda H et al (2019) Conserved fungal effector suppresses PAMP-triggered immunity by targeting plant immune kinases. *Proc Natl Acad Sci USA* 116:496–505. <https://doi.org/10.1073/pnas.1807297116>
- Kim JW, Shim SH (2019) The fungus *Colletotrichum* as a source for bioactive secondary metabolites. *Arch Pharm Res* 42:735–753. <https://doi.org/10.1007/s12272-019-01142-z>
- King BC et al (2011) Arsenal of plant cell wall degrading enzymes reflects host preference among plant pathogenic fungi. *Biotechnol Biofuels* 4:4. <https://doi.org/10.1186/1754-6834-4-4>
- Kleemann J et al (2012) Sequential delivery of host-induced virulence effectors by appressoria and intracellular hyphae of the phytopathogen *Colletotrichum higginsianum*. *Plos Pathog* 8:e1002643. <https://doi.org/10.1371/journal.ppat.1002643>
- Kloppholz S, Kuhn H, Requena N (2011) A secreted fungal effector of *Glomus intraradices* promotes symbiotic biotrophy. *Curr Biol* 21:1204–1209. <https://doi.org/10.1016/j.cub.2011.06.044>
- Kobae Y et al (2016) Phosphate treatment strongly inhibits new arbuscule development but not the maintenance of arbuscule in mycorrhizal rice roots. *Plant Physiol* 171:566–579. <https://doi.org/10.1104/pp.16.00127>

- Kubicek CP, Starr TL, Glass NL (2014) Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. *Annu Rev Phytopathol* 52(52):427–451. <https://doi.org/10.1146/annurev-phyto-102313-045831>
- Kubo Y, Furusawa I (1991) Melanin biosynthesis: prerequisite for successful invasion of the plant host by appressoria of *Colletotrichum* and *Pyricularia*. In: Cole GT, Hoch HT (eds) *The fungal spore and disease initiation in plants and animals*, New York, Plenum Publishing, pp 205–217
- Lafflamme B et al (2020) The pan-genome effector-triggered immunity landscape of a host-pathogen interaction. *Science* 367:763–768. <https://doi.org/10.1126/science.aax4079>
- Lahrman U et al (2015) Mutualistic root endophytism is not associated with the reduction of saprotrophic traits and requires a noncompromised plant innate immunity. *New Phytol* 207:841–857. <https://doi.org/10.1111/nph.13411>
- Lelwala RV et al (2019) Comparative genome analysis indicates high evolutionary potential of pathogenicity genes in *Colletotrichum tanacetii*. *PLoS ONE* 14:e0212248. <https://doi.org/10.1371/journal.pone.0212248>
- Liang X et al (2018) Pathogenic adaptations of *Colletotrichum* fungi revealed by genome wide gene family evolutionary analyses. *PLoS ONE* 13:e0196303. <https://doi.org/10.1371/journal.pone.0196303>
- Lo Presti L et al (2015) Fungal effectors and plant susceptibility. *Annu Rev Plant Biol* 66:513–545. <https://doi.org/10.1146/annurev-arplant-043014-114623>
- Lu H, Zou WX, Meng JC, Hu J, Tan RX (2000) New bioactive metabolites produced by *Colletotrichum* sp., an endophytic fungus in *Artemisia annua*. *Plant Sci* 151:67–73. [https://doi.org/10.1016/S0168-9452\(99\)00199-5](https://doi.org/10.1016/S0168-9452(99)00199-5)
- Martin F et al (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452:88–U87. <https://doi.org/10.1038/nature06556>
- Melnik RA, Hossain SS, Haney CH (2019) Convergent gain and loss of genomic islands drive lifestyle changes in plant-associated *Pseudomonas*. *ISME J* 13:1575–1588. <https://doi.org/10.1038/s41396-019-0372-5>
- Nagendran S, Hallen-Adams HE, Paper JM, Aslam N, Walton JD (2009) Reduced genomic potential for secreted plant cell-wall-degrading enzymes in the ectomycorrhizal fungus *Amanita bisporigera*, based on the secretome of *Trichoderma reesei*. *Fungal Genet Biol* 46:427–435. <https://doi.org/10.1016/j.fgb.2009.02.001>
- Nongbri PL et al (2012) Indole-3-acetaldoxime-derived compounds restrict root colonization in the beneficial interaction between arabidopsis roots and the endophyte *Piriformospora indica*. *Mol Plant Microbe Interact* 25:1186–1197. <https://doi.org/10.1094/Mpmi-03-12-0071-R>
- O'Connell RJ et al (2012) Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat Genet* 44:1060–1065. <https://doi.org/10.1038/ng.2372>
- Palmer JM et al (2013) Loss of CcIA, required for histone 3 lysine 4 methylation, decreases growth but increases secondary metabolite production in *Aspergillus fumigatus*. *PeerJ* 1:e4. <https://doi.org/10.7717/peerj.4>
- Perfect SE, Hughes HB, O'Connell RJ, Green JR (1999) *Colletotrichum*: a model genus for studies on pathology and fungal-plant interactions. *Fungal Genet Biol* 27:186–198. <https://doi.org/10.1006/fgbi.1999.1143>
- Plaumann PL, Schmidpeter J, Dahl M, Taher L, Koch CA (2018) Dispensable chromosome is required for virulence in the hemibiotrophic plant pathogen *Colletotrichum higginsianum*. *Front Microbiol* 9:1005. <https://doi.org/10.3389/fmicb.2018.01005>
- Plett JM et al (2011) A secreted effector protein of *Laccaria bicolor* is required for symbiosis development. *Curr Biol* 21:1197–1203. <https://doi.org/10.1016/j.cub.2011.05.033>
- Raffaële S, Kamoun S (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol* 10:417–430. <https://doi.org/10.1038/nrmicro2790>
- Rao S, Nandineni MR (2017) Genome sequencing and comparative genomics reveal a repertoire of putative pathogenicity genes in chilli anthracnose fungus *Colletotrichum truncatum*. *PLoS ONE* 12:e0183567. <https://doi.org/10.1371/journal.pone.0183567>



- Redman RS, Ranson JC, Rodriguez RJ (1999) Conversion of the pathogenic fungus *Colletotrichum magna* to a nonpathogenic, endophytic mutualist by gene disruption. *Mol Plant Microbe Interact* 12:969–975. <https://doi.org/10.1094/Mpmi.1999.12.11.969>
- Rodriguez RJ, White JF Jr, Arnold AE, Redman RS (2009) Fungal endophytes: diversity and functional roles. *New Phytol* 182:314–330. <https://doi.org/10.1111/j.1469-8137.2009.02773.x>
- Ryder LS, Talbot NJ (2015) Regulation of appressorium development in pathogenic fungi. *Curr Opin Plant Biol* 26:8–13. <https://doi.org/10.1016/j.pbi.2015.05.013>
- Sanchez-Vallet A et al (2018) The genome biology of effector gene evolution in filamentous plant pathogens. *Annu Rev Phytopathol* 56(56):21–40. <https://doi.org/10.1146/annurev-phyto-080516-035303>
- Shinohara Y, Kawatani M, Futamura Y, Osada H, Koyama Y (2016) An overproduction of astellolides induced by genetic disruption of chromatin-remodeling factors in *Aspergillus oryzae*. *J Antibiot* 69:4–8
- Sukno SA, Garcia VM, Shaw BD, Thon MR (2008) Root infection and systemic colonization of maize by *Colletotrichum graminicola*. *Appl Environ Microbiol* 74:823–832. <https://doi.org/10.1128/AEM.01165-07>
- Tisserant E et al (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci U S A* 110:20117–20122. <https://doi.org/10.1073/pnas.1313452110>
- Tsushima A et al (2019) Genomic plasticity mediated by transposable elements in the plant pathogenic fungus *Colletotrichum higginsianum*. *Genome Biol Evol* 11:1487–1500. <https://doi.org/10.1093/gbe/evz087>
- Tucker SL, Talbot NJ (2001) Surface attachment and pre-penetration stage development by plant pathogenic fungi. *Annu Rev Phytopathol* 39:385–417. <https://doi.org/10.1146/annurev.phyto.39.1.385>
- Varma A et al (1999) *Piriformospora indica*, a cultivable plant-growth-promoting root endophyte. *Appl Environ Microbiol* 65:2741–2744
- Wang ET et al (2012) A common signaling process that promotes mycorrhizal and oomycete colonization of plants. *Curr Biol* 22:2242–2246. <https://doi.org/10.1016/j.cub.2012.09.043>
- Xu XH et al (2014) The rice endophyte *Harpophora oryzae* genome reveals evolution from a pathogen to a mutualistic endophyte. *Sci Rep-UK* 4:5783. <https://doi.org/10.1038/srep05783>
- Yokoyama A, Izumitsu K, Irie T, Suzuki K (2019) The homeobox transcription factor CoHox1 is required for the morphogenesis of infection hyphae in host plants and pathogenicity in *Colletotrichum orbiculare*. *Mycoscience* 60:110–115. <https://doi.org/10.1016/j.myc.2018.11.001>
- Yoshino K et al (2012) Cell death of *Nicotiana benthamiana* is induced by secreted protein NIS1 of *Colletotrichum orbiculare* and is suppressed by a homologue of CgDN3. *Mol Plant Microbe Interact* 25:625–636. <https://doi.org/10.1094/Mpmi-12-11-0316>
- Yu JH, Keller N (2005) Regulation of secondary metabolism in filamentous fungi. *Annu Rev Phytopathol* 43:437–458. <https://doi.org/10.1146/annurev.phyto.43.040204.140214>
- Zou WX et al (2000) Metabolites of *Colletotrichum gloeosporioides*, an endophytic fungus in *Artemisia mongolica*. *J Nat Prod* 63:1529–1530. <https://doi.org/10.1021/np000204t>
- Zuccaro A et al (2011) Endophytic life strategies decoded by genome and transcriptome analyses of the mutualistic root symbiont *Piriformospora indica*. *Plos Pathog* 7:e1002290. <https://doi.org/10.1371/journal.ppat.1002290>

# Chapter 7

## Genome Evolution of Asexual Organisms and the Paradox of Sex in Eukaryotes



**Elvira Hörandl, Jens Bast, Alexander Brandt, Stefan Scheu, Christoph Bleidorn, Mathilde Cordellier, Minou Nowrousian, Dominik Begerow, Anja Sturm, Koen Verhoeven, Jens Boenigk, Thomas Friedl, and Micah Dunthorn**

**Abstract** The predominance of sex in eukaryotes is still enigmatic. Sex, a composed process of meiosis and mixis cycles, confers high costs but the selective advantages remain unclear. In this review, we focus on potentially detrimental effects of asexuality on genome evolution. Theory predicts that asexual lineages should suffer from lack of meiotic DNA repair, accumulation of deleterious mutations, proliferation

---

E. Hörandl (✉)

Department of Systematics, Biodiversity and Evolution of Plants (with Herbarium),  
University of Goettingen, Untere Karspüle 2, 37073 Göttingen, Germany  
e-mail: [elvira.hoerandl@biologie.uni-goettingen.de](mailto:elvira.hoerandl@biologie.uni-goettingen.de)

J. Bast

Institute of Zoology, University of Cologne, Zùlpicher Str. 47b, 50674 Köln, Germany  
e-mail: [mail@jensbast.com](mailto:mail@jensbast.com)

A. Brandt · S. Scheu

Johann Friedrich Blumenbach Institute of Zoology and Anthropology, University of Goettingen,  
Untere Karspüle 2, 37073 Göttingen, Germany  
e-mail: [alexander.brandt@biologie.uni-goettingen.de](mailto:alexander.brandt@biologie.uni-goettingen.de)

S. Scheu

e-mail: [sscheu@gwdg.de](mailto:sscheu@gwdg.de)

C. Bleidorn

Johann-Friedrich-Blumenbach Institute for Anthropology & Zoology Animal Evolution and  
Biodiversity, University of Goettingen, Untere Karspüle 2, 37073 Göttingen, Germany  
e-mail: [christoph.bleidorn@biologie.uni-goettingen.de](mailto:christoph.bleidorn@biologie.uni-goettingen.de)

M. Cordellier

Institut für Zoologie, Population genomics, University of Hamburg, Martin-Luther-King Platz 3,  
20146 Hamburg, Germany  
e-mail: [mathilde.cordellier@uni-hamburg.de](mailto:mathilde.cordellier@uni-hamburg.de)

M. Nowrousian

Fakultät für Biologie und Biotechnologie Lehrstuhl für Molekulare und Zelluläre Botanik,  
Ruhr-University Bochum, ND 7/176, Universitätsstr. 150, 44801 Bochum, Germany  
e-mail: [minou.nowrousian@ruhr-uni-bochum.de](mailto:minou.nowrousian@ruhr-uni-bochum.de)

D. Begerow

Department of Evolution of Plants and Fungi, Ruhr-University Bochum, Universitätsstraße 150,  
D-44801 Bochum, Germany  
e-mail: [dominik.begerow@rub.de](mailto:dominik.begerow@rub.de)

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_7](https://doi.org/10.1007/978-3-030-57246-4_7)

of transposable elements, among others. Here, we compare the different genomic features, life cycles, developmental pathways, and cytological mechanisms in the major eukaryotic groups, i.e., in protists, animals, fungi, and plants. In general, it is difficult to disentangle lineage-specific features from general features of asexuality. In all groups, forms of asexuality are predominantly facultative or cyclical. A variety of mixed or partial sexual developmental pathways exists, maintaining some components of sexuality, while obligate asexuality appears to be rare in eukaryotes. The strongest theoretical prediction for negative consequences of asexuality is decreased effectiveness of selection compared to sexuality. While some studies have shown increased rates of mutation accumulation in asexuals, others using whole-genome comparisons did not find this pattern. Various mechanisms exist that can alleviate the negative consequences of accumulation of negative mutations. More empirical data are needed to understand comprehensively the role of genome evolution for the maintenance of sex.

## 7.1 Introduction

It is still a core question of evolutionary biology why sexual reproduction is so predominant in eukaryotes (Otto 2009; Neiman et al. 2018). Sex can be broadly defined as “a process in which the genomes of two parents are brought together in a common cytoplasm to produce progeny which may then contain reassorted portions of the parental genomes” (Birdsell and Wills 2003). In eukaryotes, sex is a composite process consisting of meiosis, as a special form of nuclear division, and fertilization

---

A. Sturm

Institute of Mathematical Stochastics, University of Goettingen, Goldschmidtstr. 7, 37077  
Göttingen, Germany  
e-mail: [asturm@math.uni-goettingen.de](mailto:asturm@math.uni-goettingen.de)

K. Verhoeven

Department of Terrestrial Ecology, Netherlands Institute of Ecology (NIOO-KNAW),  
Droevendaalsesteeg 10, 6708 PB Wageningen, Netherlands  
e-mail: [k.verhoeven@nioo.knaw.nl](mailto:k.verhoeven@nioo.knaw.nl)

J. Boenigk

Faculty of Biology, Department of Biodiversity, University of Duisburg-Essen, Universitätsstr. 5,  
45141 Essen, Germany  
e-mail: [Jens.Boenigk@uni-due.de](mailto:Jens.Boenigk@uni-due.de)

T. Friedl

Experimentelle Phykologie und Sammlung von Algenkulturen (EPSAG), University of  
Goettingen, Nikolausberger Weg 18, 37073 Göttingen, Germany  
e-mail: [tfriedl@uni-goettingen.de](mailto:tfriedl@uni-goettingen.de)

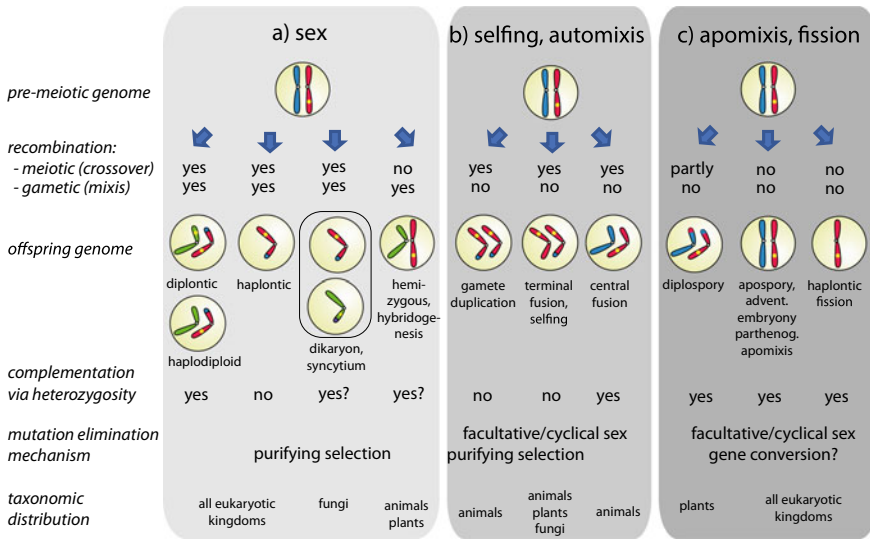
M. Dunthorn

Faculty of Biology Department of Eukaryotic Microbiology, University of Duisburg-Essen,  
Universitätsstr. 5, 45141 Essen, Germany  
e-mail: [micah.dunthorn@uni-due.de](mailto:micah.dunthorn@uni-due.de)

with fusion of nuclei (syngamy and karyogamy; Birdsell and Wills 2003; Brandeis 2018). Both components, however, imply high costs for the parental organisms: meiosis involves a breakup of favorable gene combinations and is altogether a risky, energy-consuming process. Syngamy and karyogamy (outcrossing) involve the need of two mating partners, and hence the costs of mate searching, mate finding, density dependence, eventually a cost of non-reproducing males, among others (Maynard Smith 1978; Bell 1982); these costs differ dramatically among eukaryotic taxa (Lewis 1987; Hörandl and Hadacek 2020). In general, sexual reproduction has no immediate selective advantage for the individuals performing sex. Hence, it is difficult to explain maintenance of the process and also the prevalence of obligate sexuality. Because of its high costs, sex should theoretically be replaced by asexuality (Williams 1975; Maynard Smith 1978; Bell 1982; Lewis 1987; Otto 2009). Prokaryotes demonstrate that organisms can be evolutionarily successful and diverse with asexuality, they can adapt to almost all ecological niches of this planet, and they are much older than eukaryotes—all without the burdens of meiosis–misis cycles.

Sexual reproduction probably originated already in the first eukaryotes (2010; Speijer et al. 2015; Hörandl and Speijer 2018) and had its precursors in prokaryotes (Birdsell and Wills 2003; Ramesh et al. 2005; Speijer et al. 2015; Speijer 2016). We will focus here on maintenance of sex in eukaryotes. While sexuality is still infrequent in protists, it becomes increasingly more frequent in multicellular organisms where sex and reproduction are intimately linked and reaches dominance in animals and flowering plants, with >99% of sexual species in both groups (Burt 2000). A great diversity of adaptations has evolved in plants and animals to make sexuality and mating between conspecific individuals working (Brandeis 2018), but the genetic control of meiosis (Schurko and Logsdon 2008) and the basic nuclear and cellular processes in meiosis–misis cycles have remained surprisingly conserved.

By contrast, asexuality evolved multiple times from sexual ancestors (Schwander and Crespi 2009a, b). Asexuality occurs in eukaryotes in many different cytological and developmental pathways (Fig. 7.1, Box 7.1), whereby most of them represent just various alterations of meiosis–misis cycles (Mirzaghaderi and Hörandl 2016). Sexual reproduction is hardly ever abandoned completely—in fact, “a little bit of sex,” i.e., facultative or cyclical asexuality appears to be common in fungi, plants, and animals (Simon et al. 2003; Mirzaghaderi and Hörandl 2016). The wealth of asexual cytological and reproductive pathways in eukaryotes and the nightmare of different terminologies in the traditional literature (see Schön et al. 2009) have hampered comparative empirical research (see Fig. 7.1, Box 7.1). Our knowledge on asexual genome evolution is restricted to a handful of organisms, with a strong focus on metazoans (Neiman et al. 2018). These biases made it difficult to develop generalized models of advantages of sexuality over asexuality.



**Fig. 7.1** Overview of cytological mechanisms in major sexual and asexual developmental pathways, and main consequences for genome evolution. For terminology, see Box 7.1. The yellow dots represent disadvantageous mutations that are either complemented by unmutated homologous chromosomes or are selectively eliminated by various mechanisms. **a** Sexual reproduction, with meiosis and mixis, i.e., biparental genome contributions (symbolized by green chromosome). **b** Reproduction with meiosis, but self-fertilization and hence uniparental genome contributions. **c** Reproduction without meiosis and without mixis

### Box 7.1. Forms of Asexual Reproduction and Definition of Terms

**Adventitious embryony:** a form of apomixis in plants whereby embryos develop directly out of a somatic cell.

**Apogamy:** a term for apomixis in ferns.

**Apomictic parthenogenesis:** the term for apomixis in animals. See also thelytoky.

**Apomixis:** asexual reproduction without meiosis and with parthenogenesis, usually applied to multicellular organisms. Offspring are clones of the parent.

**Apospory:** a form of apomixis in plants whereby egg cells develop from a somatic cell and develop parthenogenetically.

**Automixis:** involves meiosis and fusion of nuclei derived from the same meiosis, but nuclei originate from the same parental individual. Eggs develop parthenogenetically. Occurs mostly in animals and is usually regarded as a form of asexual reproduction.

**Central fusion:** a form of automixis involving fusion of segregated nuclei derived from meiosis I. Maintains mostly the parental heterozygosity.

**Cyclical parthenogenesis:** Life cycle comprising an alternation of sexual and asexual reproduction (many animals).

**Dikaryon:** cells that have two haploid nuclei in their major life stage (occurs in mycelia of some fungi).

**Diplospory:** a form of apomixis in plants involving a restitutional meiosis, but parthenogenetic development of egg cells.

**Facultative apomixis:** sexual and apomictic offspring are produced in parallel in one life cycle, from the same mother (frequent in plants).

**Fission:** Mitotic cell divisions. Usually applied to unicellular eukaryotes (e.g., fission yeast).

**Gamete duplication:** a form of automixis whereby chromosome sets duplicate after meiosis II. Results in complete homozygosity.

**Gynogenesis:** the male gamete is necessary for development, but is not included in genome of offspring. Offspring is formed via apomixis or automixis. Occurs in some e.g., vertebrates.

**Haplodiploidy:** Females are diploid and produce meiotically haploid eggs. Fertilized eggs become diploid females, unfertilized eggs become haploid, parthenogenetic males (occurs in insects) also called arrhenotoky. Haplodiploidy can evolve into parthenogenetic apomixis.

**Hemizygous reproduction:** the maternal genome is transmitted without recombination, while the paternal genome is only partially inherited (e.g., canina meiosis in dogroses) or not inherited (see hybridogenesis).

**Homothallic selfing:** two nuclei derived from a meiosis of the same individual (the same mating type) fuse. Occurs in fungi.

**Hybridogenesis:** a form of hemizygous reproduction whereby female genomes are inherited clonally, while male genomes are continuously recruited for one generation, but not inherited (occurs in e.g., vertebrates; see also gynogenesis).

**Parthenogenesis:** development of an egg or egg cell into an embryo without fertilization. Occurs in plants and animals.

**Pseudogamy:** a male genome contribution is needed for development of embryos, but is not inherited. Occurs in animals and plants (but with different mechanisms).

**Selfing (self-fertilization, autogamy)** involves fusion of nuclei derived from two meioses. Gametes originate from the same individual. Occurs frequently in flowering plants (also called autogamy) but is rare in animals. Selfing is in plants usually regarded a form of sexual reproduction, but cytologically more similar to automixis in animals. See also homothallic selfing.

**Sexual reproduction (amphimixis, outcrossing):** involves meiosis (consequently recombination via crossovers) and mixis (syngamy and fusion of nuclei from different individuals, and hence gametic recombination). The life cycle can be diplontic (predominant diploid stage) or haplontic (predominant haploid stage).

**Syncytium:** cells with many nuclei that may have different origin (occurs in some fungi).

Terminal fusion: a form of automixis whereby sister nuclei derived from meiosis II fuse, resulting in complete homozygosity. Occurs in animals.

Thelytoky (thelytokous parthenogenesis): females produce female offspring from unfertilized eggs. The cytological mechanism can be automictic or apomictic. Occurs in animals.

Translocation heterozygosity: a form of meiosis with ring-forming chromosomes, without recombination, combined to selfing. Occurs in a few plants.

About 20 hypotheses exist for explaining the maintenance of sex, and they can be grouped into two major theories (Williams 1975; Kondrashov 1993; Birdsell and Wills 2003): first, sex is a tool for DNA restoration to keep the integrity and functionality of the nuclear genome, whereby most mechanisms have long-term effects (Muller 1964; Bernstein et al. 1988; Holliday 1984; Hörandl 2009); second, sex and recombination create variation in the offspring for a better short-term response and adaptive potential to environmental variability (Weismann 1904; Maynard Smith 1978).

Long-term benefits of sex act at the lineage level and unfold only after many successive generations. They derive, for example, from the ability to effectively combine beneficial alleles arising in different lineages (Fisher–Muller hypothesis; Fisher 1930; Muller 1932); from restoration of genotypes least loaded with deleterious mutations that are continuously lost by drift in finite populations (Muller's ratchet; Muller 1964), from removal of linkage between beneficial and deleterious alleles (Hill and Robertson 1966), or from enabling repair of deleterious alleles via meiotic gene conversion (Bernstein et al. 1988; Marais 2003). Hypotheses on the selective advantage of sex inherently predict that asexuality comes with specific disadvantages that manifest themselves in the genome over evolutionary time scales.

Short-term benefits of sex unfold under strong fluctuating directional selection at the level of individuals and genes. Furthermore, individual selection allows sexual populations to withstand succumbing to the cost of sex, e.g., via quick replacement by asexual offshoots. The rationale is that sex enables effective selection by generating variation in fitness through break-up of linked loci with opposite fitness effects (Felsenstein 1974). Such negative linkage disequilibria can be generated via combined effects of drift and selection in finite populations (Hill–Robertson interference; Hill and Robertson 1966). Thus, for sex to be beneficial in the short term, the trajectory of selection must change constantly (Brooks 1988); e.g., through coevolving parasites (fluctuating epistasis, Red-Queen; reviewed in Jaenike 1978; Hamilton 1980) or spatial variation in the availability of resources coupled with abiotic conditions (Scheu and Drossel 2007; Song et al. 2011). A constant change of selection scenarios, however, is an unrealistic assumption, and hence an insufficient explanation for the maintenance of obligate sexuality. Theories and empirical studies on the maintenance of sex have been recently reviewed by Jalvingh et al. (2016) and Neiman et al. (2018).

These two theories are not mutually exclusive, and also pluralistic approaches combining both ideas have been proposed (West et al., 1999; Neiman et al. 2017). Here, we will focus on theories regarding genome evolution, because the universality of the genetic code is comparable among all eukaryotes. Moreover, methodological advances of genome sequencing in the last two decades have brought upon new empirical data for genome evolution in eukaryotes and have shed a new light on old theories. First, we will review the most important theories about the role of sex in maintaining genomic integrity; second, we will report the state of the art on asexual genome evolution over all major groups of eukaryotes, including protists, algae, animals, fungi, and plants. Finally, we will draw preliminary conclusions and outline perspectives for future research.

## 7.2 The Theoretical Background for Asexual Genome Evolution

Several authors have argued that a major function of sex is maintaining genomic integrity (Muller 1964; Holliday 1984; Bernstein et al. 1988; summarized in Birdsell and Wills 2003; Hörandl 2009). Sex as tool for DNA restoration involves three different components: first, DNA repair, such as repair of physical DNA damage on the molecular structure of DNA (Bernstein and Bernstein 1991); second, elimination of deleterious mutations, i.e., changes in the sequence of base pairs); and third, maintenance of DNA methylation. These three components may act in combination (Hörandl 2009).

### 7.2.1 DNA Repair

The need for DNA repair is a constant and immediate pressure for all organisms. Transformation, a process through which prokaryotes are able to uptake exogenous DNA, is a probable precursor of sex. In prokaryotes, this mechanism is used for physical DNA repair, as prokaryotes needed to cope with DNA breaks caused by UV irradiation, temperature extremes, and other sources of oxidative damage (Birdsell and Wills 2003). In eukaryotes, many enzymes for homologous recombinational repair evolved into meiosis genes or became integrated in the eukaryotic meiosis machinery (Ramesh et al. 2005; Malik et al. 2008; Schurko and Logsdon 2008). Bernstein et al. (1988) and Bernstein and Bernstein (1991) recognized that the cytological processes during prophase I of meiosis are mainly directed to repair DNA double-strand breaks, while recombination resulting from cross-overs appears to be an infrequent by-product of this process. Meiosis research has so far confirmed these findings (reviewed by Mirzaghaderi and Hörandl 2016). Eukaryotic cells may have



evolved advanced DNA repair mechanisms because of increased intracellular oxidative stress caused by aerobic respiration and other redox-active metabolic processes that have increased damage of DNA. The evolution of linear chromosomes, of the nuclear envelope and of homologous recombinational repair at meiosis I, could have been driven by this selective force (Speijer et al. 2015; Hörandl and Speijer 2018). Homologous recombination is the most accurate and least mutagenic repair mechanism of DNA breaks (Bleuyard et al. 2006), but it requires a second homologous chromosome as a template. This requirement enforced a diploid stage in the life cycle and a chromosome from another individual with a different history of damage. The requirement of diploidy could have been the major driver for the evolution of mixis (Hörandl and Speijer 2018). In modern meiosis, minor DNA lesions, i.e., DNA radicals caused by oxidative damage, could initiate meiotic homologous recombinational DNA repair (Hörandl and Hadacek 2013).

### 7.2.2 *Mutation Accumulation*

Inaccurate repair of DNA damage is a major source of mutagenesis (Friedberg et al. 2006). Non-homologous repair mechanisms are prone to mutagenesis, while homologous recombinational repair, acting during meiosis, is the least mutagenic mechanism (Bleuyard et al. 2006). Most mutations are either neutral or have negative effects, and deleterious mutations are a constant threat for functionality of genomes. Mutations (changes of the sequence of DNA bases) cannot be actively repaired; they can only be eliminated or favored by selection according to their effect on the fitness of the organism (Bernstein et al. 1988). The theory of Muller's ratchet predicts that without sex and recombination, deleterious mutations would accumulate in a lineage over generations in finite populations (Muller 1964; Kondrashov 1988; Charlesworth et al. 1993a, b). Recombination can reinstall non-mutated genomes in the offspring, whereas in an asexual lineage, the class of non-loaded or least-loaded class of individuals will get lost by drift. This irreversible process represents a click of the ratchet. Hence, in asexual lineages, mutations will accumulate gradually over generations until a certain threshold of deleterious mutations is reached when the lineage goes extinct.

Muller's ratchet clicks at a constant rate depending on population size ( $N$ ), the deleterious mutation rate per haploid genome ( $U$ ), and the strength of selection ( $s$ ) (e.g., Jain 2008). Infinite populations are not affected by Muller's ratchet as a class of non-mutated individuals will always be present, while in finite populations the least-loaded class will be more easily lost by drift (Birdsell and Wills 2003). The deleterious mutation rate of course differs between eukaryotic organisms and has been empirically tested for just a few model organisms. For diploid or polyploid organisms, the mutation rate  $U$  multiplies with the ploidy level  $c$  (Gerstein and Otto 2009). In diploid or polyploid genomes, or in fungal dikaryons, mutations appear in the heterozygous state, and functional alleles may "mask" the recessive mutated gene copy from selection (Kondrashov and Crow 1991; Gerstein and Otto 2009).

Hence, mutation accumulation should have little effect in the short term, but may be in the long term more severe in polyploids than in diploids, because mutations are eliminated slower than in diploid lineages (Gerstein and Otto 2009). The actual speed of the ratchet depends further on recombination rates (Charlesworth et al. 1993a, b), epistasis between genes (Kondrashov 1988), and effects of beneficial mutations (Muller 1932; Crow and Kimura 1965). Residual sexuality with a little bit of recombination is sufficient to halt Muller's ratchet (Green and Noakes 1995; Hodac et al. 2019). The deterministic mutation model by Kondrashov (1988) involves synergistic epistasis between deleterious mutations such that their combined effect is more severe than the sum of their individual effects. Truncating selection is assumed to act on individuals that carry many such synergistic mutations, and hence the death of these individuals will eliminate many mutations from the population. Sex is here of great advantage as recombination breaks up linkage disequilibrium, i.e., the negative gene combinations, and increases the variance on which selection can act upon. However, empirical evidence rather suggested that negative epistasis between deleterious mutations is uncommon (Kouyos et al. 2007). Epistasis appears under the precondition of pleiotropy and evolves in a dynamic manner, depending on robustness and complexity of genomes (de Visser and Elena 2007; de Visser et al. 2011). Finally, the rare beneficial mutations have to be considered as large sexual populations can incorporate them more rapidly than asexual ones, whereas in small populations, there would be no difference between sexual and asexual populations in the speed of incorporating beneficial mutations (Fisher 1930; Muller 1932).

For prokaryotes, mutation accumulation via Muller's ratchet is a lesser problem because of large population size, rapid generation turnover, and small, haploid genomes. Prokaryotes further avoid mutation accumulation via frequent lateral gene transfer from one lineage to the other, which is under strong purifying selection (Vos et al. 2015). In eukaryotes, however, the parameters determining genome evolution become more diverse and more complex. The diversity of life cycles, mutation rates, population sizes, the more complex organization of genomes with manifold more genes, and the diversity of reproductive systems with varying recombination rates makes it difficult to apply one model that fits all organisms. While theoretical studies support the idea that mutation elimination is a strong advantage of sex, empirical research on various organisms gives an equivocal picture (see Sects. 7.3–7.6).

### 7.2.3 *Epigenetic Damage and Transposable Elements*

Holliday (1984) suggested that removal of epigenetic defects in germline cells and reinstallation of lost cytosine methylations would be a major function of meiosis. Similar as for the DNA repair hypothesis, an integer template of a second homologous chromosome is needed for maintenance methylation. Methylations are crucial for cell differentiation, development, and regulation of gene expression and are evolutionary ancient in eukaryotes (Law and Jacobsen 2010). Although DNA methylations are to

some extent heritable (Law and Jacobsen 2010), there still is uncertainty about long-term evolutionary effects. In the context of paradox of sex theories, little attention has been paid to Holliday's theory. Most research on differences of methylation patterns between sexual and asexual organisms has been done in plants (see Sect. 7.6).

Another aspect of genome evolution related to methylation is the proliferation of transposable elements (Hickey 1992). Three hypotheses exist regarding TEs in eukaryotic genomes: (1) TEs could be sexually transmitted and spread like "genomic parasites," and so their spread could be avoided by asexuality (Wright and Finnegan 2001). (2) Purifying selection after meiotic recombination can act against TE proliferation in sexual species. Hence, asexual species may suffer from an uncontrolled proliferation of TEs and may be even driven to extinction (Arkhipova and Meselson 2000, 2005). (3) Finally, there might be no relationship between TE proliferation and mode of reproduction. We will also review new insights into this complex topic in the next paragraphs.

### 7.3 Asexual Genome Evolution in Protists

The diversity of protists is extremely large, i.e., it covers all major eukaryotic clades, except the animals and fungi (Amorphea clade) and plants (Embryophyta clade) (Boenigk et al. 2015; Adl et al. 2019). Although once thought to be proto-animals or proto-plants, protists are now known to form numerous independent lineages that comprise the bulk of the genetic and metabolic diversity within the eukaryotes (Keeling and Burki 2019). Since the origins of eukaryotes sometime in the Proterozoic (Eme et al. 2014), protists have evolved into numerous ecological roles that are central in most ecosystems (Azam et al. 1983; Weisse et al. 2016).

#### 7.3.1 *Sex and Reproduction in Protists*

Although meiotic sex likely originated in the common ancestor of all extant eukaryotes, many protist species and higher clades were long thought to be asexual because of the absence of direct observations of mating between individuals (Schurko et al. 2009; Dunthorn and Katz 2010). On theoretical grounds based on the genetic advantages of recombination, it has been argued that most protists are nevertheless likely cryptically sexual at least occasionally (Dunthorn and Katz 2010; Hofstatter and Lahr 2019). And, on experimental grounds, evidence for this cryptic sex throughout the protists has been found by inventorying meiotic genes in different putative asexual lineages (Ramesh et al. 2005; Malik et al. 2008; Chi et al. 2014b; Dunthorn et al. 2017; Hofstatter et al. 2018; Kraus et al. 2018), although these meiotic genes could be used just for selfing or for non-canonical genetic pathways (Dunthorn et al. 2017). Sex has, however, been lost in some lineages, some of which could be ancient (Doerder 2014).

Meiotic sex in protists was elegantly shown to reverse the effects of Muller's ratchet by Calkins (1919), in what was called "one of the most important experimental results in biology" (Bell 1988). After preventing mating in the ciliate *Uroleptus*, division rates slowed down. Presumably, the slowdown in division rates was due to Muller's ratchet. But after mating, faster division rates were restored and thus, sex was able to "rejuvenate" cultures of ciliates in the laboratory.

Although most protists are likely meiotically sexual and therefore can use sexual recombination to reverse Muller's ratchet, the majority of reproduction in the largely unicellular protists is just mitotic cell divisions (Dunthorn and Katz 2010). Given their small sizes and fast mitotic division times, protists therefore can have massive population sizes (Finlay 2002), although the effective population sizes are smaller (Watts et al. 2013). These massive population sizes could potentially prevent Muller's ratchet by allowing selection to be more powerful than drift, as the rate of loss of few mutations depends on the absolute number of individuals (Bell 1988). Within most natural communities in different environments, though, not all protist species will have these massive population sizes (Dunthorn et al. 2014; Logares et al. 2014), thus the rare species may not necessarily be able to prevent or slow down Muller's ratchet.

### 7.3.2 Protistan Genome Structures and Muller's Ratchet

Genome architecture is known to help drive the strength and direction of evolution (Lynch 2007). There are three aspects to genome structure in protists that may be able to prevent, or at least slow down, Muller's ratchet.

The first aspect is that many protists are highly polyploid (Raikov 1982). For example, even between closely related ciliate species in *Paramecium*, there are multiple rounds of whole-genome duplications (Aury et al. 2006), and extreme levels of gene repetitions are observed in the foraminiferan *Reticulomyxa filosa* (Glöckner et al. 2014). Maciver (2016) suggested that in protists such high levels of polyploidy may help prevent or slow down Muller's ratchet, if deleterious mutations in one of the gene copies are replaced by other copies. This mechanism could be powerful if there is pervasive gene conversion and as long as there is always a gene copy with fewer deleterious mutations. Independent of the evolutionary time or the number of generations, gene conversion would not be successful if all copies had deleterious mutations. The extent of high levels of ploidy level is unknown within most protist species and higher clades, so it is unknown whether polyploidy may prevent or slow down the effects of Muller's ratchet and in how many species such effects occur.

The second aspect is the highly dynamic nuclear genome sizes of many protists (Parfrey et al. 2008). The genome sizes can greatly increase and decrease throughout the life cycle. In the well-known *Amoeba proteus*, this excessive DNA is eliminated late in interphase before mitotic division. Chromatin is ejected during this process

from the nucleus into the cytoplasm, where it is presumably degraded and recycled. Goodkov et al. (2019) suggested that such DNA extrusions may allow protists to prevent or slow down Muller's ratchet, if gene copies with deleterious mutations are being eliminated. This mechanism could be powerful if deleterious mutations are selectively removed, but if removal is random then the most abundant copies will just likely be eliminated. And as with polyploidy above, the extent of elimination of excessive DNA (and the excretion of deleterious gene copies) is unknown within most protist species and higher clades.

The third aspect only occurs in ciliates having two types of nuclei in each cell: micronuclei and macronuclei (Katz 2001; Lynn 2008). Micronuclei are transcriptionally inactive and are involved in sex and the formation of macronuclei. Division of micronuclei occurs through canonical mitosis or meiosis where homologous chromosomes are segregated by a spindle apparatus, although a functional synaptonemal complex is missing (Chi et al. 2014a). Macronuclei are transcriptionally active and are highly polyploid. Division of macronuclei occurs through amitosis, where chromosomes are randomly distributed without a spindle apparatus (Morgens and Cavalcanti 2015; Zhang et al. 2019). This random distribution of chromosomes during amitosis can lead to the loss of gene and chromosome copies in resulting progeny and eventual death if all copies are lost (Bell 1988; Zhang et al. 2019), which is a form of deleterious mutation accumulation. The ciliate *Tetrahymena thermophila* somehow controls chromosome copy number during amitosis, although the mechanism is not clear (Zhang et al. 2019), and some type of similar mechanism is likely found in many more ciliate groups (Morgens and Cavalcanti 2015). If ciliates go through meiosis and then self, however, any gene or chromosome loss, and consequently Muller's ratchet, can be reversed because new macronuclei with the full complement of genes and chromosomes are newly formed by the new micronuclei after selfing unless there were also gene losses in the micronuclei (Bell 1988).

## 7.4 Asexual Genome Evolution in Animals

Obligate asexuality (here female-producing parthenogenesis; thelytoky) is assumed to be rare in animals, found in approximately 0.1% of species (White 1977; Bell 1982). However, this number is based on the very scarce occurrence or even absence (e.g., birds and mammals) of asexuality in vertebrates (White 1977). Recent quantitative studies indicate that obligate asexuality has evolved much more frequently in species-rich non-vertebrate taxa like arthropods and molluscs, for example, with up to 1.5% in haplodiploid arthropod species under conservative estimates (under more relaxed assumptions up to 38% (van der Kooi et al. 2017)). Thus, the occurrence of obligate asexuality in animals seems vastly underestimated and understudied.

The number of parthenogenetic species largely depends on the rate with which incipient asexual lineages are generated and subsequently lost again in an animal group, but very little is known about the frequency of such transitions (Schwander and Crespi 2009a, b). The transition to asexuality from sexual progenitors can be caused

by different mechanisms, such as hybridization, endosymbiont infection, and spontaneous mutations (for an overview see Neiman et al. 2014; Jaron et al. 2019). Moreover, offspring can be generated from unfertilized eggs via many different cellular mechanisms, such as apomixis and automixis with a plethora of diverse subforms (Fig. 7.1; see e.g. Suomalainen et al. 1987; Schön et al. 2009). The underlying mechanisms for both the transition to and cytology of asexuality can have profound and different consequences for genome evolution (Engelstädter 2017; Jaron et al. 2019; Parker et al. 2019), but are, as yet, little studied in animals. By contrast, explaining the short-term and long-term benefits of sex has received considerable attention in both theoretical and empirical studies on animals (Sharp and Otto 2016; Neiman et al. 2017, 2018).

### 7.4.1 Accumulation of Slightly Deleterious Mutations

The Hill–Robertson effect and Muller’s ratchet predict a reduction of the effectiveness of purifying selection resulting in the accumulation of fixed and segregating slightly deleterious mutations in asexual species (see Introduction and, e.g., Keightley and Otto 2006). This prediction received equivocal support in animals: four out of eight available single gene-based studies found less effective purifying selection in asexual as compared to closely related sexual species (for details see Hartfield 2016; Glémin et al. 2019). However, a number of genome-based studies found excessive among gene variation in effectiveness of purifying selection indicating that interpreting single gene-based results as representative for the genome level is problematic (see Neiman et al. 2018). Further, only one (*Timema* stick insects; Bast et al. 2018) out of nine genome-based studies supports less effective purifying selection in asexuals compared to their sexual sister species (Ament-Velásquez et al. 2016; Bast et al. 2018; Brandt et al. 2017, 2019; Kraaijeveld et al. 2016; Lindsey et al. 2018; Ollivier et al. 2012; Tucker et al. 2013; Warren et al. 2018). Notably, two out of these studies based on genomic data even showed increased effectiveness of purifying selection in asexuals, including ancient asexual oribatid mites, contrary to predictions (Kraaijeveld et al. 2016; Brandt et al. 2017).

What factors can alleviate the predicted negative effects allowing asexuals to escape mutational meltdown? Large population sizes have been discussed as an important factor maintaining effective purifying selection under asexuality (Gordo and Charlesworth 2000; Rice and Friberg 2009; Normark and Johnson 2011; Ross et al. 2013). Many widely distributed and small-bodied animals have potentially very large populations (Gaston et al. 1997; White et al. 2007). Indeed, census population densities of very old asexual taxa (e.g., the above-mentioned oribatid mites) can exceed  $10^5$  individuals per square meter and generally feature larger population sizes than their sexual relatives (Maraun et al. 2012). In addition to population sizes, extensive DNA repair and/or homogenizing processes like mitotic gene conversion, or facultative recombination during cyclical parthenogenesis may play an important

role by removing deleterious alleles and exposing recessive deleterious mutations to selection (Charlesworth et al. 1993a, b; Marais 2003).

### 7.4.2 *Accumulation of Deleterious Transposable Elements in Animals*

In non-recombining genome regions of sexual species, deleterious transposable elements (TEs) can rapidly and substantially increase in numbers (e.g., *Drosophila* neo-Y chromosomes; Bachtrog et al. 2008). A number of empirical studies tested whether such accumulation of TEs extends to the genome scale in completely non-recombining genomes of obligate asexual animals, potentially generating selection for sex at the lineage level (similar to the accumulation of point mutations, but possibly more rapidly). No overall genomic difference could be detected between asexual and related sexual animals, only very variable and lineage-specific TE dynamics were found (Bast et al. 2016; Szitenberg et al. 2016; Jiang et al. 2017; Jaron et al. 2019). This lack of difference is likely due to a number of confounding factors not related to reproductive modes (such as, e.g., hybridization and polyploidization) that can affect TE dynamics (Arkhipova and Rodriguez 2013). Despite no overall difference, higher TE turnover in cyclically sexual *Daphnia pulex* indicates that sex facilitates both the spread and elimination of TEs (Jiang et al. 2017). The few investigated older asexual animals harbor few and inactive TEs (Flot et al. 2013; Bast et al. 2016). Whether this stems from the evolution of benign TEs via suppression mechanisms (as indicated in experimentally evolved yeast; Bast et al. 2019; for a review on mechanisms see Koonin et al. 2020) or from the immediate extinction of asexual lineages with high TE contents after the loss of sex remains an open question.

### 7.4.3 *The “Meselson Effect”*

Homologous chromosomes in asexual organisms are expected to accumulate mutations independently of each other in regions sheltered from loss of heterozygosity and diverge in parallel. This should lead to high levels of heterozygosity and parallel topological resemblance of haplotype subtrees over populations (Birky 1996; Judson and Normark 1996; Mark Welch and Meselson 2000). Testing this “Meselson effect” is important because its presence is regarded as strong support for the complete absence of sex and theoretically opens the possibility for dating the transition to asexuality in the absence of fossils (Normark et al. 2003). As yet, only single gene-based studies in asexual *Timema* stick insects and fissiparous *Dugesia* flatworms have shown the expected haplotype divergence pattern (Schwander et al. 2011; Leria et al. 2019). Large within-individual variance levels were found in a number of different

invertebrates, e.g., the apomictic *Meloidogyne* root-knot nematodes and the ribbon worm *Lineus pseudolacteus* but (later) attributed to divergence between homeologs derived from hybridization (Lunt 2008; Ament-Velásquez et al. 2016; Jaron et al. 2019). Similarly, large within-individual variance in bdelloid rotifer species has been shown to result from an ancient genome duplication event resulting in tetraploidy and reflect divergence between ancient homologs (so-called ohnologs) instead of haplotypes (Mark Welch et al. 2008; Flot et al. 2013; Nowell et al. 2018). In other animal species, which show no sign of the Meselson effect, such as darwinulid ostracods or tramini aphids, haplotype divergence has been putatively reduced due to homogenizing processes like mitotic gene conversion (Normark 1999; Schön and Martens 2003). A genome-wide comparison of asexual and sexual lineages of *Daphnia pulex* showed that loss of heterozygosity via such homogenizing processes is a dramatically more powerful force than accumulation of new mutations (Tucker et al. 2013).

#### **7.4.4 Genomic Features Based on Single Asexual Genome Studies**

The genomes of singular asexual animal species featured some peculiarities that were suggested to be generally linked to asexuality, such as horizontal gene transfer (HGT), genomic rearrangements, gene family expansions, gene losses, and gene conversion (Danchin et al. 2010; Flot et al. 2013; Faddeeva-Vakhrusheva et al. 2017; for a full review see Jaron et al. 2019). Many of these features are related to the idea that contrary to sexually reproducing organisms, asexuals do not require chromosomal homolog pairing during meiosis, which potentially leads to increased fixation of structural variants. However, none of the features were systematically replicated across 26 published animal genomes, suggesting that these genomic peculiarities are lineage-specific and not generally linked to asexuality (Jaron et al. 2019). Testing this idea further needs whole-genome studies on structural variants in asexuals compared to closely related sexual species (see outlook).

#### **7.4.5 Evolutionary Scandals: Ancient Asexuals**

Genomic consequences of asexuality with detrimental fitness effects are expected to accumulate over time and eventually drive asexual lineages to extinction (Gabriel et al. 1993; Lynch et al. 1993). However, few asexual lineages have persisted and even diversified in the absence of sex for considerable periods of time (Judson and Normark 1996; Schön et al. 2009; Schwander et al. 2011). The most notorious examples include bdelloid rotifers, darwinulid ostracods, and several parthenogenetic taxa of oribatid mites (Judson and Normark 1996; Schön et al. 2009). Among these, bdelloid rotifers have, so far, received most attention (Mark Welch and Meselson 2000; Flot et al.



2013). Recent studies, however, have indicated that cryptic gene exchange renders them quasi-sexual (Signorovitch et al. 2015; Debortoli et al. 2016; Schwander 2016; Vakhrusheva et al. 2018; Laine et al. 2020). The amount and mechanism of cryptic sex and DNA uptake remain controversial (Flot et al. 2018; Wilson et al. 2018). Data on genome evolution in asexual oribatid mites and darwinulid ostracods are scarce. While for asexual oribatid mites two studies showed effective purifying selection and decreased load of transposable elements (Bast et al. 2016; Brandt et al. 2017), there are currently no genome data-based studies in darwinulid ostracods. More studies on these two animal groups are urgently required as truly ancient asexual lineages are invaluable for generating insights into the long-term selective advantage of sex.

## 7.5 Asexual Genome Evolution in Fungi

Fungi are an ancient, species-rich lineage of eukaryotes with a wide variety of lifestyles (Hawksworth and Lücking 2017; Spatafora et al. 2017). Fungi can be unicellular (yeasts) or multicellular (filamentous fungi); the latter forming cell filaments (hyphae) that form tissue-like networks (mycelia). Fungi can undergo asexual propagation either through mitotic cell division in the case of yeasts, or through hyphal fragmentation or the formation of mitotic spores in the case of filamentous fungi (Golan and Pringle 2017). Sexual propagation in fungi, leading to the formation of meiotic spores, is usually induced under species-specific conditions and can be facultative or an integral part of the life cycle as is the case for a number of plant-pathogenic fungi (Bennett and Turgeon 2016; Peraza-Reyes and Malagnac 2016; Coelho et al. 2017; Lee and Idnurm 2017). While many fungal species were described as asexual for decades, genome analyses as well as population genomics studies and crossing experiments in the laboratory have led to the discovery of sexual propagation in many presumed asexual fungal species (Dyer and Kück 2017). Therefore, it is currently not known whether any truly asexual fungal lineages exist that have completely lost the ability to undergo sexual propagation. In the following sections, we will discuss current knowledge in the two largest fungal groups, the Ascomycota and Basidiomycota, and then briefly mention some recent results for the Glomeromycotina, a group of plant symbionts that have been discussed as a long-term asexual group of fungi. We will finish the review with some thoughts on the continuum of sexual versus asexual propagation in fungi.

### 7.5.1 *Modes of Reproduction in Ascomycota*

Ascomycota is named after their sexual sporangia (asci, singular ascus), and sexual propagation has been studied in great detail at the molecular level in a number of ascomycete model organisms (Bennett and Turgeon 2016; Zickler and Espagne 2016; Pöggeler et al. 2018). Especially the genes required for mating and meiosis

are well known in ascomycetes and can be used as molecular markers for the presence of cryptic sexual development in species where sexual propagation has not been observed yet. However, it has to be noted that meiotic genes may have functions outside of sexual reproduction, e.g., in stress-related ploidy changes as was recently shown in the human pathogenic fungus *Cryptococcus neoformans* (Zhao et al. 2020). Therefore, crossing experiments or population genetic studies are important to study actual sexual reproduction as described below. Mating in fungi is genetically regulated by so-called mating type (*MAT*) loci that contain at least one *MAT* gene (Kües et al. 2011; Bennett and Turgeon 2016). In ascomycetes, *MAT* genes often encode transcription factors that regulate downstream genes required for sexual reproduction. In self-sterile (heterothallic) ascomycete species, successful mating is only possible between partners with compatible *MAT* loci, whereas self-fertile (homothallic) species often encode compatible *MAT* genes within one genome (Heitman 2015; Pöggeler et al. 2018).

In the last century, fungal species for which no sexual stage was known from nature or laboratory observations were designated as “deuteromycetes” or “imperfect fungi,” and it was assumed that such species had lost the capacity to undergo sexual reproduction. Within the ascomycetes, this applied to up to 40% of surveyed taxa (Dyer and Kück 2017). However, population studies starting in the 1990s indicated that cryptic sex can exist in such species. The first study to show this analyzed polymorphic genetic markers in clinical isolates of *Coccidioides immitis*, the causal agent of the valley fever. Marker distribution in isolates was consistent with genetic recombination as opposed to clonal propagation of this fungus (Burt et al. 1996). Another line of evidence came after the first genomes of supposedly asexual species were sequenced in the early 2000s, and *MAT* genes as well as meiosis-specific genes were found to be present and to not have accumulated mutations (Pöggeler 2002; Galagan et al. 2005). A major breakthrough was achieved when it was shown that natural isolates of the supposedly asexual species *Aspergillus fumigatus* can undergo sexual development in the laboratory (O’Gorman et al. 2009). Since then, sexual reproduction under laboratory conditions was demonstrated for a number of supposedly asexual ascomycetes, and it is currently not clear if any truly asexual lineages can exist in the long term.

### 7.5.2 Modes of Reproduction in Basidiomycota

Similar to the Ascomycota, the Basidiomycota are named after their meiosporangium, the basidium, which in contrast to the ascus bears its spores externally instead of internal spore development. However, the life cycle is very similar with a sexual phase bearing meiotic spores and an asexual phase giving rise to millions of mitotic conidia in many lineages. As in Ascomycota, sexual structures are often unknown or overlooked, as they can be reduced to a few cells only being microscopically visible (Sampaio 2004; Oberwinkler 2017). Several lineages have managed to link developmental stages like parasitism or vector-based dispersal to the alternating

life cycle (Morrow and Fraser 2009). Sexual compatibility is usually mediated by two mating loci, one of which is coding for a pheromone/pheromone receptor system controlling syngamy, while the second is coding for homeodomain (HD) transcription factors relevant for maintenance of the dikaryon, regular cell divisions, and filamentous growth (Raudaskoski and Kothe 2010). The separation of the two mating loci on two chromosomes leads to a tetrapolar mating system in most Basidiomycota, with multiple alleles of the various genes in several lineages (Kües et al. 2011).

As the haploid phase of the life cycle is often characterized by a saprobic, yeast-like stage, most Basidiomycota from early diverging lineages are isolated as haploid cultures from nature and their sexual structures are unknown or at least not observed in culture. Genera like *Pseudozyma*, *Rhodospodium*, *Tilletiopsis*, or *Cryptococcus* were used to describe these so-called asexual species. Phylogenetic studies revealed that these genera are polyphyletic and mixed with sexual species suggesting overlooked sexual stages in some lineages (Begerow et al. 2000). However, several lineages like the genera *Malassezia*, *Moniliella*, *Tilletiopsis washingtonensis* s.l. seem to be completely asexual (without signs of sexual stages), although the mating genes seem to be present as in the case of *Malassezia* (Wang et al. 2015; Saunders et al. 2012). Many studies focused on mating under laboratory conditions to identify sexual structures. These studies could identify several mechanisms to maintain a sexual life cycle even without a compatible mating partner (Lin and Heitmann 2007, David-Palma et al. 2016) and such pseudo-sexual strategies might be common among Basidiomycota in several lineages (Coehlo et al. 2017). Functions of pheromones and pheromone receptors might be thus very diverse including functions not involved in mating and reproduction. Recently, it was shown that non-mating-type-specific receptors are common in Agaricomycetes (Kües et al. 2011), and therefore, functions of predicted pheromone receptors in potentially asexual lineages need to be elucidated to allow conclusive remarks on the presence or absence of sexual stages. At present, it is not yet clear if obligate asexuals exist among the Basidiomycota.

### 7.5.3 *Glomeromycotina: Ancient Asexuals or Cryptic Sex?*

A case in point about the difficulty of identifying truly asexual fungal species might be the Glomeromycotina. They belong to the Mucoromycota, a sister group to ascomycetes and basidiomycetes (Spatafora et al. 2017). The Glomeromycotina are mostly obligate plant symbionts that form the widespread arbuscular mycorrhiza with the roots of land plants (Lanfranco et al. 2016). Despite their environmental ubiquity, cultivation of Glomeromycotina in the laboratory is difficult due to their metabolic dependence on the host plant, and no sexual stages have been observed in nature, probably because their life is spent completely underground and thus they are difficult to observe in their natural environment. Glomeromycotina were considered as ancient asexuals; however, targeted searches for meiotic genes as well as genomic studies confirmed the presence of meiotic genes and putative *MAT* genes in Glomeromycotina genomes, making it likely that a sexual cycle exists in these

species (Halary et al. 2011; Tisserant et al. 2013; Lin et al. 2014; Ropars et al. 2016). Therefore, based on the available data, the Glomeromycotina cannot be described as ancient asexuals.

#### 7.5.4 Other Reproductive Strategies Influencing Genome Evolution

Given recent genomic insights in the distribution of sex-related genes and sexual propagation in fungi, it has been suggested that many fungal species might be considered not as completely sexual or asexual, but rather as consisting of isolates on a continuum of sexual reproduction ranging from fully fertile to asexual (Dyer and Kück 2017). This raises the question under what conditions a sexual or an asexual lifestyle might be advantageous specifically for a fungal species or isolate. One possible factor might be the degree of ploidy. Even though some fungi are diploids, the majority of ascomycetes and basidiomycetes harbor haploid nuclei. However, in filamentous fungi, the mycelia are usually coenocytic with two or more nuclei sharing a common cytoplasm (Maheshwari 2005). Furthermore, fungi can undergo vegetative hyphal fusions between different individuals of the same species leading to exchange of genetically different nuclei (Daskalov et al. 2017). Thus, deleterious mutations in one nucleus might be masked by functional copies in other nuclei, making the mycelia functionally similar to the cells of heterozygous diploid organisms (Fig. 7.1). However, especially hyphal fusion and subsequent nuclear exchange come with the risk of spreading, for example, infectious agents or transposons. Therefore, many fungi have evolved heterokaryon incompatibility systems that allow vegetative hyphal fusion only between compatible partners. Calculations based on allele frequencies for different incompatibility systems in the model fungus *Neurospora crassa* suggest that the likelihood for compatible interactions between germinating vegetative spores is rather low (Gonçalves et al. 2019). Thus, it is possible that at least in *N. crassa* propagation is biased toward sexual propagation (during which the vegetative incompatibility systems are turned off), because sexual propagation is limited to dedicated partitions of the mycelium, thereby preventing spreading of infectious agents to the rest of the mycelium.

Another point to be considered with respect to the advantages of sexual or asexual reproduction is the spreading of transposable elements (TEs), which in fungi could in principle occur during sexual propagation or during the above-mentioned vegetative hyphal fusion. It is interesting to note, though, that many sequenced fungal genomes have a low TE content compared to other eukaryotic genomes (Castanera et al. 2016; Spatafora et al. 2017; Stajich 2017). Studies in several fungal model organisms have revealed that at least five different genome defense systems evolved within the fungi that protect organisms from the spread of TEs and other repeats (Gladyshev 2017). In *N. crassa* alone, three genome defense mechanisms are known, one of which is active during vegetative growth, a second in the dikaryotic phase directly prior

to karyogamy, and the third during meiosis (Shiu et al. 2001; Gladyshev 2017). Thus, it appears that at least *N. crassa* has every contingency covered and is genetically prepared to counter transposon spread during sexual and asexual propagation. Genome defense in other fungi is less well studied, but it seems likely that genome defense mechanisms against transposable elements are present in some form in other species as well.

Population studies were performed for a few species only and therefore data on the relevance of recombination in fungi are broadly lacking. However, homothallism has been discussed as a common strategy to perform selfing and maintain at least parts of sexual recombination. For example, *Cryptococcus neoformans* is known to perform a unisexual or pseudo-sexual life cycle in addition to a classical sexual cycle (Lin et al. 2005; Ni et al. 2013). Thus, fungi display a huge variety of mixed forms between truly sexual and asexual species highlighting their great potential to adapt to diverse needs of reproduction. Obviously, facultative asexuality is here predominant as well as in other eukaryotes. A black-and-white system of sex/asex does not exist.

## 7.6 Asexual Genome Evolution and Epigenomics in Plants

### 7.6.1 Asexual Reproduction in Plants

Asexual reproduction occurs in land plants mostly in ferns, in a form called apogamy (Grusz 2016), and in flowering plants as apomixis, the asexual reproduction via seeds (Asker and Jerling 1992; Mogie 1992, see Box 1 for terminology). Land plants have a life cycle of alternating diplontic sporophytes (producing meiospores) and a haplontic gametophyte (producing gametes). Asexual reproduction keeps this life cycle but avoids meiosis–mixis cycles in many different ways. We will focus here on flowering plants. Apomixis is in angiosperms scattered across the phylogeny and occurs in about 2% of genera, with many different developmental pathways (Hojsgaard et al. 2014; Fig. 7.1). Studies on asexual genome evolution are scarce due to practical difficulties: first, plant genomes are complex and can vary dramatically in size (Michael 2014). Angiosperms have undergone ancient and recent genome duplications, resulting in gene duplications and diversification of gene functions (Jiao et al. 2011; Leebens-Mack et al. 2019). Second, asexuality does not occur in model organisms like *Arabidopsis* or in major crops plants. For this reason, genomic resources are also scarce. The only completely sequenced reference genomes for gametophytic apomixis is published from *Boechea*, a relative of *Arabidopsis* (Kantama et al. 2007; Kliver et al. 2018), and for sporophytic apomixis in *Citrus* (Wang et al. 2017). (Gametophytic and sporophytic apomixis are characterized by embryo development either from cells in the megagametophyte or directly from somatic cells in the sporophyte, respectively). Third, most research has focused so far on understanding genetic control mechanisms of apomixis rather than on evolutionary questions, with the major aim to introduce apomixis into crops (Ozias-Akins and Conner 2019).

Apomixis is heritable (Ozias-Akins and van Dijk 2007), but regulatory mechanisms turned out to be unexpectedly complex and rely mostly on differential expression of many genes that regulate the sexual pathway (Koltunow and Grossniklaus 2003; Sharbel et al. 2010; Hand and Koltunow 2014; Schmidt et al. 2014). Apomixis is usually facultative and occurs mostly in polyploids or in diploid hybrids. Recent studies on natural apomicts, however, suggest that apomixis emerges already at the diploid level and is then directly and indirectly established by polyploidy (Hojsgaard and Hörandl 2019). Facultative sexuality can also involve selfing (fertilization of egg cells with pollen from the same plant) as most apomicts are self-compatible (Hörandl 2010). Selfing is an otherwise common sexual pathway in plants (Schemske and Lande 1985), resulting in an increase of homozygosity and loss of genotypic diversity in the offspring. However, little is known about frequencies and effects of selfing in otherwise facultative apomictic lineages.

### 7.6.2 Case Studies on the Possible Effects of Muller's Ratchet in Plants

The first comparative study on asexual genome evolution used transcriptomes of flowering buds in the *Ranunculus auricomus* complex (Pellino et al. 2013). This system comprised obligately diploid sexual and facultative apomictic hexaploids. Transcriptome analysis by using dN/dS ratios revealed that both sexual and asexual genomes are under purifying selection without signs of genome-wide accumulation of deleterious mutations as evolutionary theory would predict (see above). The outlier genes with elevated non-synonymous to synonymous (dN/dS) ratios in the sexual/asexual comparisons belonged to genes involved in sporogenesis and gametogenesis, and hence may relate to functional aspects of apomixis rather than to mutation accumulation. However, the lack of a related reference genome hampered a comprehensive gene annotation. Nevertheless, signatures of allelic sequence divergence were detected in the hexaploid apomictic genomes, probably due to hybrid origin. The same system was studied using a mathematical model, incorporating empirical data on the degree of facultative recombination and different selection scenarios (Hodac et al. 2019). Results confirmed the hypothesis that even a low degree of facultative sexuality in these hexaploid apomictic lineages was sufficient to counteract Muller's ratchet. Purifying selection might be specifically efficient in the meiotically formed, haplontic gametophytes, a stage in which many genes are expressed, and hence deleterious mutations can be efficiently eliminated.

A similar study was performed on transcriptomes on four sexual/asexual species pairs of the genus *Oenothera* (Hollister et al. 2015). In this genus, asexual reproduction occurs in the quite unusual form of permanent translocation heterozygosity, i.e., a reciprocal translocation of chromosomes that results in a ring formation at meiosis such that chromosomes pair only at their tips. This ring form of meiosis results in complete suppression of meiotic recombination and segregation. Seeds are

usually formed via selfing; i.e., the parental genotype is maintained. Hollister et al. (2015) found indeed elevated levels of heterozygosity and increased accumulation of non-synonymous mutations in the asexual lineages compared to sexual species. This system supports the hypothesis that obligate asexuality results in mutation accumulation. However, also rare (facultative) outcrossing has been found in *O. biennis* (Maron et al. 2018), which might counteract Muller's ratchet.

A recent study using genome sequences on sexual/apomictic, diploid species pairs of *Boechera* revealed high levels of heterozygosity in apomicts, mostly due to hybrid origin (Lovell et al. 2017). Analysis of mutation accumulation was performed at different types of different genomic sites: (1) conserved non-coding sites, (2) conserved coding sites, (3) sites, where any mutation causes an amino acid substitution, and (4) sites where any mutation is synonymous. Mutation accumulation in asexuals was found to be significantly higher in categories (1) and (3), but not in (2), indicating that purifying selection is still present, but more relaxed in phylogenetically derived sites. Mutation accumulation was found to be independent from hybrid origin, although it is difficult to entangle contemporary mutations from the ancestral ones in the conspecific hybrid. The authors did not consider effects of facultative apomixis, although variable proportions of sexual/asexual seeds occurred in 11 of their 13 apomictic samples (Lovell et al. 2017). Facultative apomixis occurs also in other taxa of *Boechera* (Aliyu et al. 2010). Specific studies on evolution of RNA helicases in sexual and apomictic *Boechera* revealed that mutation accumulation is further depending on gene function (Kiefer et al. 2020).

Taken together, the presence of Muller's ratchet was overall confirmed in plants, but a little bit of sex seems to be sufficient to counteract the accumulation of deleterious mutations. The predominance of facultative sexuality and lack of ancient asexuals in plants fit to this scenario. The degree of facultative sexuality, however, is highly flexible in plants and can be influenced positively by environmental stress conditions (Klatt et al. 2016, 2018; Ulum et al. 2020). Such a stress response of the reproductive mode is ploidy-dependent (Ulum et al. 2020). The possible combination of apomixis to selfing adds another level of complexity to understand mutation dynamics in asexual plant lineages. These dynamics, however, have not yet been investigated.

### 7.6.3 Studies on the Epigenome and Transposable Elements

Plant mode of reproduction can have consequences for the proper functioning of the epigenetic mechanisms that suppress TE activity. Epigenetic mechanisms, and specifically DNA methylation, silence TEs, can modulate gene expression. In plants, DNA (cytosine) methylation occurs in all sequence contexts (CG, CHG, and CHH, where H = C, T or A). The enzymatic pathways for depositing and maintaining methylation marks differ between contexts, as do their functions and dynamics. Broadly speaking, DNA methylation in plant gene bodies often occurs in the CG

context, which shows strong transgenerational stability but unclear functional relation to gene expression (Wendte et al. 2019). TEs can be densely methylated in all sequence contexts, which is under active control by a small RNA-guided DNA methylation mechanism (RdDM, RNA-directed DNA methylation) and which is associated with transcriptional silencing (Slotkin and Martienssen 2007; Matzke et al. 2015). The pathways that lead to DNA methylation silencing of TEs involve small RNAs that are used to guide methylation to specific TE loci (Matzke et al. 2015).

Because large portions of plant genomes can consist of TEs (for instance, up to 85% in the maize genome; Schnable et al. 2009), the epigenetic mechanisms that protect the genome from their uncontrolled transposition have potentially large consequences for genome evolution. In recent years, asexual plants have become popular model systems to study the causes and evolutionary consequences of DNA methylation variation (Richards et al. 2017), because (1) epigenetic effects are more easily detected in uniform genomic backgrounds, and (2) epigenetic variation is hypothesized to be of comparably high relevance for adaptation in asexuals that show little DNA sequence variation (Verhoeven and Preite 2014).

A relevant proximate question related to the impact of epigenetic mechanisms on asexual genome evolution is: How does the absence of meiosis in asexuals affect the stability and genomic patterns of epigenetic variation, and what consequences does this have for genome evolution? While detailed analysis in asexual plants is limited by the availability of high-quality reference genomes (Richards et al. 2017), relevant insights come from understanding the epigenetic processes that take place during sexual plant reproduction. In comparison to mammals, which undergo extensive DNA methylation erasure during gametogenesis and early embryogenesis (Feng et al. 2010), plants experience relatively limited DNA methylation resetting between generations. DNA methylation in CG contexts in particular shows high transgenerational stability (Johannes et al. 2009). However, DNA methylation that is under control of the RdDM pathway shows interesting dynamics: in both male and female germlines, cells that accompany the germ cells, but not the germ cells themselves, undergo active demethylation (Ibarra et al. 2012); the endosperm shows reduced DNA methylation levels, and the developing embryo shows a gradual increase in non-CG methylation (Bouyer et al. 2017).

In pollen, the loss of DNA methylation in the vegetative cell (which does not contribute genetic information to the next generation) releases TE activity, which results in TE-derived expressed transcripts that are subsequently degraded into small RNAs. It has been shown that these small RNAs are transported to the sperm cells (which do contribute to the next generation) where they can be used by the RdDM machinery to target corresponding TE sequences for silencing (Martinez et al. 2016). A similar process seems to occur in female plant germ lines (Ibarra et al. 2012). It is believed that this mechanism functions to reinforce TE silencing in the germ cells and the resulting embryo (Slotkin et al. 2009).

Asexual reproduction that does not involve the above germline epigenetic processes may therefore result in less efficient epigenetic silencing of TEs. For instance, lack of sex-reinforced TE silencing is thought to underpin an abnormal fruit phenotype (“mantled” fruit that produces less oil) that arose in oil palms under



tissue culture. This stable phenotypic variant is caused by loss of methylation in a LINE retrotransposon (Ong-Abdullah et al. 2015). Beyond variable silencing of TEs, we can speculate that compromised silencing results in increased transposition rates over longer evolutionary time scales, thereby contributing to mutational degeneration of asexual lineages.

## 7.7 Conclusion and Outlook

It has become clear that most of the classical predictions on the disadvantages of asexual reproduction need (re-)evaluation on a whole-genome basis. Moreover, for testing most hypothesis on the maintenance of sex, it is imperative to compare replicates of independently derived asexual lineages to closely related sexual species at both the population level and species level to disentangle true consequences of asexuality from confounding lineage-specific patterns. Such comparative studies are needed for a phylogenetically broad representation of the eukaryote tree of life (Burki et al. 2019).

The classical prediction that sex is imperative for effective purging of deleterious mutations is not universally met. It remains an open question whether large population sizes and/or effective repair mechanisms facilitate effective selection in ancient asexuals. Importantly, if strategies exist that avoid long-term detrimental effects of asexuality, research efforts should focus more on the immediate, short-term benefits of sex that require identification of eco-evolutionary dynamics. Facultative asexuality might occur much more often than expected in animals and fungi, which might alter the perception of costs and benefits of sex. More knowledge is required on the frequency of transitions to asexuality in natural populations as well as identifying its underlying molecular mechanisms and cytology.

The diversity of different developmental pathways and genomic features of asexual eukaryotes needs to be considered for empirical genome studies. The prevalence of mixed systems (such as facultative, cyclical, or intermittent sexuality also often combined to selfing) complicates predictive models. Why are not all eukaryotes capable of such mixed systems with a “little bit of sex,” which would preserve both variability and maintenance of favorable genotypes, both in the short and long term? Perhaps we should ask the question why and how sex has ever become obligate? Here, a better understanding of regulatory mechanisms and functions of meiosis–mixis cycles will be essential. Genomic studies beyond mutation screenings, targeting DNA repair and mutagenesis as well as epigenetic effects and TE dynamics, are needed to understand the actual advantages of obligate sexuality compared to facultative or obligate asexuality.

**Acknowledgements** We thank the editors, Marie-Hélène Rome and Pierre Pontarotti, for inviting us to contribute to this book, and one anonymous referee for valuable comments on the manuscript. The work on this chapter was supported by Deutsche Forschungsgemeinschaft (DFG), projects 4395/4-1 and 4395/10-1 to E.H., BA 5800/3-1 to J.B., DU1319/5-1 to M.D., NO407/7-1 to M.N.

## References

- Adl SM, Bass D, Lane CE, Lukes J, Schoch CL, Smirnov A et al (2019) Revisions to the classification, nomenclature, and diversity of eukaryotes. *J Eukaryotic Microbiol* 66(1):4–119. <https://doi.org/10.1111/jeu.12691>
- Aliyu OM, Schranz ME, Sharbel TF (2010) Quantitative variation for apomictic reproduction in the genus *Boechera* (Brassicaceae). *Am J Bot* 97(10):1719–1731. <https://doi.org/10.3732/ajb.100188>
- Ament-Velázquez SL, Figuet E, Ballenghien M, Zattara EE, Norenburg JL, Fernández-Álvarez FA et al (2016) Population genomics of sexual and asexual lineages in fissiparous ribbon worms (Lineus, Nemertea): hybridization, polyploidy and the Meselson effect. *Mol Ecol* 25:3356–3369
- Arkhipova I, Meselson M (2000) Transposable elements in sexual and ancient asexual taxa. *Proc Natl Acad Sci USA* 97(26):14473–14477. <https://doi.org/10.1073/pnas.97.26.14473>
- Arkhipova I, Meselson M (2005) Deleterious transposable elements and the extinction of asexuals. *BioEssays* 27:76–85
- Arkhipova IR, Rodriguez F (2013) Genetic and epigenetic changes involving (retro)transposons in animal hybrids and polyploids. *Cytogenet Genome Res* 140:295–311
- Asker S, Jerling L (1992) Apomixis in plants. CRC Press, Boca Raton
- Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM et al (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171–178
- Azam F, Fenchel T, Field JG, Gray JS, Meyerreil LA, Thingstad F (1983) The ecological role of water-column microbes in the sea. *Mar Ecol Prog Ser* 10:257–263
- Bachtrog D, Hom E, Wong KM, Maside X, de Jong P (2008) Genomic degradation of a young Y chromosome in *Drosophila miranda*. *Genome Biol.* 9:R30
- Bast J, Schaefer I, Schwander T, Maraun M, Scheu S, Kraaijeveld K (2016) No accumulation of transposable elements in asexual arthropods. *Mol Biol Evol* 33:697–706
- Bast J, Parker DJ, Dumas Z, Jalvingh KM, Van Tran P, Jaron KS et al (2018) Consequences of asexuality in natural populations: insights from stick insects. *Mol Biol Evol* 35:1668–1677
- Bast J, Jaron KS, Schuseil D, Roze D, Schwander T (2019) Asexual reproduction reduces transposable element load in experimental yeast populations. *Elife* 8:e48548
- Begerow D, Bauer R, Boekhout T (2000) Phylogenetic placements of ustilaginomycetous anamorphs as deduced from nuclear LSU rDNA sequences. *Mycol Res* 104:53–60
- Bell G (1982) The masterpiece of nature: the evolution and genetics of sexuality. California Press, Berkeley
- Bell G (1988) Sex and death in protozoa: the history of an obsession. Cambridge University Press, Cambridge
- Bennett RJ, Turgeon BG (2016) Fungal sex: the ascomycota. In: Heitman J, Howlett BJ, Crous PW, Stukenbrock EH, James TY, Gow NAR (eds) The fungal kingdom. American Society for Microbiology. <https://doi.org/10.1128/microbiolspec.FUNK-0005-2016>
- Bernstein C, Bernstein H (1991) Aging, sex and DNA repair. Academic Press, San Diego
- Bernstein H, Byerly H, Hopf F, Michod RE (1988) Is meiotic recombination an adaptation for repairing DNA, producing genetic variation, or both? In: Michod RE, Levin BR (eds) The evolution of sex. Sinauer Ass Inc., Sunderland, pp 139–160
- Birdsell JA, Wills C (2003) The evolutionary origin and maintenance of sexual recombination: a review of contemporary models. In: Macintyre RJ, Clegg MT (eds) Evolutionary biology. Springer, Boston, pp 27–138
- Birky CW Jr (1996) Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* 144:427–437
- Bleuyard JY, Gallego ME, White CI (2006) Recent advances in understanding of the DNA double-strand break repair machinery of plants. *DNA Repair* 5(1):1–12. <https://doi.org/10.1016/j.dnarep.2005.08.017>
- Boenigk J, Wodniok S, Glücksman E (2015) Biodiversity and earth history. Springer, Berlin

- Bouyer D, Kramdi A, Kassam M, Heese M, Schnittger A, Roudier F et al (2017) DNA methylation dynamics during early plant life. *Genome Biol* 18. <https://doi.org/10.1186/s13059-017-1313-0>
- Brandeis M (2018) New-age ideas about age-old sex: separating meiosis from mating could solve a century-old conundrum. *Biological Rev* 93(2):801–810. <https://doi.org/10.1111/brv.12367>
- Brandt A, Schaefer I, Glanz J, Schwander T, Maraun M, Scheu S et al (2017) Effective purifying selection in ancient asexual oribatid mites. *Nat Commun* 8:873
- Brandt A, Bast J, Scheu S, Meusemann K, Donath A, Schütte K, Machida R, Kraaijeveld K (2019) No signal of deleterious mutation accumulation in conserved gene sequences of extant asexual hexapods. *Sci Rep* 9:5338
- Brooks LA (1988) The evolution of recombination rates. In: Michod, RE, Levin, BR (eds) *The evolution of sex*. Sinauer, Sunderland
- Burki F, Roger AJ, Matthew W, Brown MW, Simpson AGB (2019) The new tree of eukaryotes. *Trends Ecol Evol*. <https://doi.org/10.1016/j.tree.2019.08.008>
- Burt A (2000) Perspective: sex, recombination, and the efficacy of selection—was Weismann right? *Evolution* 54(2):337–351
- Burt A, Carter DA, Koenig GL, White TJ, Taylor JW (1996) Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proc Nat Acad Sci USA* 93(2):770–773
- Calkins GN (1919) *Uroleptus mobilis* Engelm. II. Renewal of vitality through conjugation. *J Exp Zool* 29:121–156
- Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J et al (2016) Transposable elements versus the fungal genome: impact on whole-genome architecture and transcriptional profiles. *PLOS Genet* 12:e1006108
- Cavalier-Smith T (2010) Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. *Biol Direct* 5:7
- Charlesworth D, Morgan M, Charlesworth B (1993a) Mutation accumulation in finite outbreeding and inbreeding populations. *Genetics Res* 61(01):39–56
- Charlesworth D, Morgan MT, Charlesworth B (1993b) Mutation accumulation in finite populations. *J Hered* 84:321–325
- Chi J, Mahé F, Loidl J, Logsdon J, Dunthorn M (2014a) Meiosis gene inventory of four ciliates reveals the prevalence of a synaptonemal complex-independent crossover pathway. *Mol Biol Evol* 31:660–672
- Chi J, Parrow MW, Dunthorn M (2014b) Cryptic sex in *Symbiodinium* (Alveolata, Dinoflagellata) is supported by an inventory of meiotic genes. *J Eukaryot Microbiol* 61:322–327
- Coelho MA, Bakkeren G, Sun S, Hood ME, Giraud T (2017) Fungal sex: the Basidiomycota. In: Heitman J, Howlett BJ, Crous PW, Stukenbrock EH, James TY, Gow NAR (eds) *The fungal kingdom*. American Society for Microbiology. <https://doi.org/10.1128/microbiolspec.FUNK-0046-2016>
- Crow JF, Kimura M (1965) Evolution in sexual and asexual populations. *Amer Naturalist* 99:439–450
- Danchin EGJ, Rosso M-N, Vieira P, de Almeida-Engler J, Coutinho PM, Henrissat B, Abad P (2010) Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proc Nat Acad Sci USA* 107:17651–17656
- Daskalov A, Heller J, Herzog S, Fleißner A, Glass NL (2017) Molecular mechanisms regulating cell fusion and heterokaryon formation in filamentous fungi. *Microbiol Spectr* 5:FUNK-0015-2016
- David-Palma M, Sampaio JP, Goncalves P (2016) Genetic dissection of sexual reproduction in a primary homothallic basidiomycete. *PLoS Genet* 12(6):e1006110. <https://doi.org/10.1371/journal.pgen.1006110>
- de Visser JA, Elena SF (2007) The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat Rev Genet* 8:139–149
- de Visser JA, Cooper TF, Elena SF (2011) The causes of epistasis. *Proc Biol Sci* 278:3617–3624
- Debertoli N, Li X, Eyres I, Fontaneto D, Hespels B, Tang CQ et al (2016) Genetic exchange among bdelloid rotifers is more likely due to horizontal gene transfer than to meiotic sex. *Curr Biol* 26:723–732

- Doerder FP (2014) Abandoning sex: multiple origins of asexuality in the ciliate Tetrahymena. *BMC Evol Biol* 14:112
- Dunthorn M, Katz LA (2010) Secretive ciliates and putative asexuality in microbial eukaryotes. *Trends Microbiol* 18:183–188
- Dunthorn M, Stoeck T, Clamp J, Warren A, Mahé F (2014) Ciliates and the rare biosphere: a review. *J Eukaryot Microbiol* 61:404–409
- Dunthorn M, Zufall RA, Chi J, Paszkiewicz K, Moore K, Mahé F (2017) Meiotic genes in colpodean ciliates support secretive sexuality. *Genome Biol Evol* 9:1781–1787
- Dyer PS, Kück U (2017) Sex and the imperfect fungi. In: Heitman J, Howlett BJ, Crous PW, Stukenbrock EH, James TY and Gow NAR (eds) *The fungal kingdom*. American Society for Microbiology, <https://doi.org/10.1128/microbiolspec.FUNK-0043-2017>
- Eme L, Sharpe SC, Brown MW, Roger AJ (2014) On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb Perspect Biol* 6:a016139
- Engelstädter J (2017) Asexual but not clonal: evolutionary processes in automictic populations. *Genetics* 206:993–1009
- Faddeeva-Vakhrusheva A, Kraaijeveld K, Derks MFL, Anvar SY, Agamennone V, Suring W et al (2017) Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genomics* 18:493
- Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78:737–756
- Feng SH, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. *Science* 330(6004):622–627. <https://doi.org/10.1126/science.1190614>
- Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science* 296:1061–1063
- Fisher RA (1930) *The genetical theory of natural selection*. Oxford University Press, Oxford
- Flot J-F, Hespels B, Li X, Noel B, Arkhipova I, Danchin EGJ, Hejnol A, Henrissat B, Koszul R, Aury J-M et al (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* 500:453–457
- Flot JF, Debortoli N, Hallet B, Narayan J (2018) Reply to cross-contamination explains “inter and intraspecific horizontal genetic transfers” between asexual bdelloid rotifers (Wilson, Nowell & Barraclough 2018) *BioRxiv*. <https://www.biorxiv.org/content/10.1101/368209v1.abstract>
- Friedberg EC, Wlaker GC, Siede W, Wood RD, Schultz RA, Ellenberger T (2006) *DNA repair and mutagenesis*, 2 edn. American Society for Microbiology, Washington DC
- Gabriel W, Lynch M, Bürger R (1993) Muller’s ratchet and mutational meltdowns. *Evolution* 47:1744–1757
- Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S et al (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 438(7071):1105–1115
- Gaston KJ, Blackburn TM, Lawton JH (1997) Interspecific abundance-range size relationships: an appraisal of mechanisms. *J Anim Ecol* 66:579–601
- Gerstein AC, Otto SP (2009) Ploidy and the causes of genomic evolution. *J Hered* 100(5):571–581. <https://doi.org/10.1093/jhered/esp057>
- Gladyshev E (2017) Repeat-induced point mutation and other genome defense mechanisms in fungi. In: Heitman J, Howlett BJ, Crous PW, Stukenbrock EH, James TY, Gow NAR (eds) *The fungal kingdom*. American Society for Microbiology
- Glémin S, François CM, Galtier N (2019) Genome evolution in outcrossing vs. selfing vs. asexual species. *Methods Mol Biol* 1910:331–369
- Glöckner G, Hülsmann N, Schleicher M, Noegel AA, Eichinger L, Gallinger C et al (2014) The genome of the foraminiferan *Reticulomyxa filosa*. *Curr Biol* 24:11–18. <https://doi.org/10.1016/j.cub.2013.11.027>
- Golan JJ, Pringle A (2017) Long-distance dispersal of fungi. *Microbiol Spectr* 5:FUNK-0047-2016
- Gonçalves AP, Heller J, Span EA, Rosenfield G, Do HP, Palma-Guerrero J et al (2019) Allorecognition upon fungal cell-cell contact determines social cooperation and impacts the acquisition of multicellularity. *Curr Biol* 29:3006–3017

- Goodkov AV, Berdieva MA, Podlipaeva YI, Demin SYu (2019) The chromatin extrusion phenomenon in *Amoeba proteus* cell cycle. *J Eukaryot Microbiol.* <https://doi.org/10.1111/jeu.12771>
- Gordo I, Charlesworth B (2000) The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics* 154:1379–1387
- Green RF, Noakes DLG (1995) Is a little bit of sex as good as a lot? *J Theor Biol* 174(1):87–96. <https://doi.org/10.1006/jtbi.1995.0081>
- Gruz AL (2016) A current perspective on apomixis in ferns. *J Syst Evol* 54(6):656–665. <https://doi.org/10.1111/jse.12228>
- Halary S, Malik SB, Lildhar L, Slamovits CH, Hijri M, Corradi N (2011) Conserved meiotic machinery in *Glomus* spp., a putatively ancient asexual fungal lineage. *Genome Biol Evol* 3:950–958
- Hamilton WD (1980) Sex versus non-sex versus parasite. *Oikos* 35:282–290
- Hand ML, Koltunow AMG (2014) The genetic control of apomixis: asexual seed formation. *Genetics* 197(2):441–450. <https://doi.org/10.1534/genetics.114.163105>
- Hartfield M (2016) Evolutionary genetic consequences of facultative sex and outcrossing. *J Evol Biol* 29:5–22
- Hawksworth DL, Lücking R (2017) Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiol Spectr* 5:FUNK-0052-2016
- Heitman J (2015) Evolution of sexual reproduction: a view from the fungal kingdom supports an evolutionary epoch with sex before sexes. *Fungal Biol Reviews* 29:108–117
- Hickey DA (1992) Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes. *Genetica* 86:269–274
- Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8:269–294
- Hodac L, Klatt S, Hojsgaard D, Sharbel T, Hörandl E (2019) A little bit of sex prevents mutation accumulation even in apomictic polyploid plants. *BMC Evol Biol* 19:170. <https://doi.org/10.1186/s12862-019-1495-z>
- Hofstatter PG, Lahr DJG (2019) All eukaryotes are sexual, unless proven otherwise. *BioEssays* 41:e1800246. <https://doi.org/10.1002/bies.201800246>
- Hofstatter PG, Brown MW, Lahr DJG (2018) Comparative genomics supports sex and meiosis in diverse Amoebozoa. *Mol Biol Evol* 10:3118–3128. <https://doi.org/10.1093/gbe/evy241>
- Hojsgaard D, Hörandl E (2019) The rise of apomixis in natural plant populations. *Front Plant Sci* 10. <https://doi.org/10.3389/fpls.2019.00358>
- Hojsgaard D, Klatt S, Baier R, Carman JG, Hörandl E (2014) Taxonomy and biogeography of apomixis in angiosperms and associated biodiversity characteristics. *Crit Rev Plant Sci* 33(5):414–427. <https://doi.org/10.1080/07352689.2014.898488>
- Holliday R (1984) The biological significance of meiosis. *Symposia Soc ExperBiol* 38:381–394
- Hollister JD, Greiner S, Wang W, Wang J, Zhang Y, Wong GK-S et al (2015) Recurrent loss of sex is associated with accumulation of deleterious mutations in *Oenothera*. *Mol Biol Evol* 32(4):896–905
- Hörandl E (2009) A combinatorial theory for maintenance of sex. *Heredity* 103(6):445–457. <https://doi.org/10.1038/hdy.2009.85>
- Hörandl E (2010) The evolution of self-fertility in apomictic plants. *Sex Pl Repr* 23(1):73–86. <https://doi.org/10.1007/s00497-009-0122-3>
- Hörandl E, Hadacek F (2013) The oxidative damage initiation hypothesis for meiosis. *Sex Pl Repr* 26:351–367
- Hörandl E, Hadacek F (2020) Oxygen, life forms, and the evolution of sexes in multicellular eukaryotes. *Heredity* (in press). <https://doi.org/10.1038/s41437-020-0317-9>
- Hörandl E, Speijer D (2018) How oxygen gave rise to eukaryotic sex. *Proc B-Biol Sci* 285(1872):20172706. <https://doi.org/10.1098/rspb.2017.2706>

- Ibarra CA, Feng XQ, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA et al (2012) Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* 337(6100):1360–1364. <https://doi.org/10.1126/science.1224839>
- Jaenike J (1978) An hypothesis to account for the maintenance of sex within populations. *Evol Theory* 3:191–194
- Jain K (2008) Loss of least-loaded class in asexual populations due to drift and epistasis. *Genetics* 179(4):2125–2134
- Jalvingh K, Bast J, Schwander T (2016) Sex, evolution and maintenance of. *Encyclopedia of evolutionary biology* [Internet], pp 89–97. Available from: <https://doi.org/10.1016/b978-0-12-800049-6.00144-x>
- Jaron KS, Bast J, Nowell RW, Rhyker Ranallo-Benavidez T, Robinson-Rechavi M, Schwander T (2019) Genomic features of asexual animals. *bioRxiv* [Internet], p 497495. Available from: <https://www.biorxiv.org/content/10.1101/497495v2>
- Jiang X, Tang H, Ye Z, Lynch M (2017) Insertion polymorphisms of mobile genetic elements in sexual and asexual populations of *Daphnia pulex*. *Genome Biol Evol* 9:362–374
- Jiao Y, Wickett NJ, Ayyampalayam S, Chanderali AS, Landherr L, Ralph PE et al (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97–100
- Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N et al (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *Plos Genetics* 5(6). <https://doi.org/10.1371/journal.pgen.1000530>
- Judson OP, Normark BB (1996) Ancient asexual scandals. *Trends Ecol Evol* 11:41–46
- Kantama L, Sharbel TF, Schranz ME, Mitchell-Olds T, de Vries S, de Jong H (2007) Diploid apomicts of the *Boechera holboellii* complex display large-scale chromosome substitutions and aberrant chromosomes. *Proc Nat Acad Sci USA* 104(35):14026–14031. <https://doi.org/10.1073/pnas.0706647104>
- Katz LA (2001) Evolution of nuclear dualism in ciliates: a reanalysis in light of recent molecular data. *Int J Syst Evol Microbiol* 51:1587–1592
- Keeling P, Burki F (2019) Progress towards the tree of eukaryotes. *Curr Biol* 29:R808–R817
- Keightley PD, Otto SP (2006) Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 443:89–92
- Kiefer M, Nauerth BH, Volkert C, Ibberson D, Loreth A, Schmidt A (2020) Gene function rather than reproductive mode drives the evolution of RNA helicases in sexual and apomictic *Boechera*. *Genome Biol Evol* <https://doi.org/10.1093/gbe/evaa078>
- Klatt S, Hadacek F, Hodač L, Brinkmann G, Eilerts M, Hojsgaard D et al (2016) Photoperiod extension enhances sexual megaspore formation and triggers metabolic reprogramming in facultative apomictic *Ranunculus auricomus*. *Front Plant Sci* 7:278. <https://doi.org/10.3389/fpls.2016.00278>
- Klatt S, Schinkel CC, Kirchheimer B, Dullinger S, Hörandl E (2018) Effects of cold treatments on fitness and mode of reproduction in the diploid and polyploid alpine plant *Ranunculus kuepferi* (*Ranunculaceae*) *Ann Bot* 121(7):1287–1298
- Kliver S, Rayko M, Komissarov A, Bakin E, Zhermakova D, Prasad K et al (2018) Assembly of the *Boechera retrofracta* genome and evolutionary analysis of apomixis-associated genes. *Genes* 9(4):16. <https://doi.org/10.3390/genes9040185>
- Koltunow AM, Grossniklaus U (2003) Apomixis: a developmental perspective. *Ann Rev Plant Biol* 54:547–574. <https://doi.org/10.1146/annurev.arplant.54.110901.160842>
- Kondrashov AS (1988) Deleterious mutations and the evolution of sexual reproduction. *Nature* 336(6198):435–440. <https://doi.org/10.1038/336435a0>
- Kondrashov AS (1993) Classification of hypotheses on the advantage of amphimixis. *J Hered* 84:372–387
- Kondrashov AS, Crow JF (1991) Haploidy or diploidy: which is better? *Nature* 351(6324):314–315
- Koonin EV, Makarova KS, Wolf YI, Krupovic M (2020) Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet* 21:119–131
- Kouyos RD, Silander OK, Bonhoeffer S (2007) Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol Evol* 22(6):308–315

- Kraaijeveld K, Anvar SY, Frank J, Schmitz A, Bast J, Wilbrandt J et al (2016) Decay of sexual trait genes in an asexual parasitoid wasp. *Genome Biol Evol* 8:3685–3695
- Kraus D, Chi J, Boenigk J, Beisser D, Graupner N, Dunthorn M (2018) Putatively asexual chrysophytes have meiotic genes: evidence from transcriptomic data. *Peer J* 6:e5894
- Kües U, James TY, Heitman J (2011) Mating type in basidiomycetes: unipolar, bipolar, and tetrapolar patterns of sexuality. In: Pöggeler S, Wöstemeyer J (eds) *The Mycota XIV. Evolution of fungi and fungal-like organisms*. Springer, Berlin, pp 97–160
- Laine V, Sackton T, Meselson M (2020) Sexual reproduction in bdelloid rotifers. *bioRxiv*. <https://doi.org/10.1101/2020.08.06.23959>
- Lanfranco L, Bonfante P, Genre A (2016) The mutualistic interaction between plants and arbuscular mycorrhizal fungi. *Microbiol Spectr* 4:FUNK-0012-2016
- Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204–220. <https://doi.org/10.1038/nrg2719>
- Lee SC, Idnurm A (2017) Fungal sex: the Mucoromycota. In: Heitman J, Howlett BJ, Crous PW, Stukenbrock EH, James TY, Gow NAR (eds) *The fungal kingdom*. American Society for Microbiology. <https://doi.org/10.1128/microbiolspec.FUNK-0041-2017>
- Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW et al (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574(7780):679. <https://doi.org/10.1038/s41586-019-1693-2>
- Leria L, Vila-Farré M, Solà E, Riutort M (2019) Outstanding intraindividual genetic diversity in fissiparous planarians (*Dugesia*, Platyhelminthes) with facultative sex. *BMC Evol Biol* 19:130
- Lewis WM (1987) The cost of sex. In: Stearns SC (ed) *The evolution of sex and its consequences*. Birkhäuser, Basel, pp 33–57
- Lin X, Hull CM, Heitman J (2005) Sexual reproduction between partners of the same mating type in *Cryptococcus neoformans*. *Nature* 434(7036):1017–1021. <https://doi.org/10.1038/nature03448>
- Lin X, Heitman J (2007) Mechanisms of homothallism in fungi and transitions between heterothallism and homothallism, pp 35–57. In: Heitman J, Kronstad JW, Taylor JW, Casselton LA (eds) *Sex in fungi: molecular determination and evolutionary implications*. ASM Press, Washington, DC. <https://doi.org/10.1128/9781555815837.ch3>
- Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D et al (2014) Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLOS Genet* 10(1):e1004078. <https://doi.org/10.1371/journal.pgen.1004078>
- Lindsey ARI, Kelkar YD, Wu X, Sun D, Martinson EO, Yan Z, Rugman-Jones PF, Hughes DST, Murali SC, Qu J, Dugan S, Lee SL, Chao H, Dinh H, Han Y, Doddapaneni HV, Worley KC, Muzny DM, Ye G, Gibbs RA, Richards S, Yi SV, Stouthamer R, Werren JH (2018) Comparative genomics of the miniature wasp and pest control agent *Trichogramma pretiosum*. *BMC Biol* 16:54
- Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R et al (2014) Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol* 24:813–821
- Lovell JT, Williamson RJ, Wright SI, McKay JK, Sharbel TF (2017) Mutation accumulation in an asexual relative of *Arabidopsis*. *Plos Genet* 13(1). <https://doi.org/10.1371/journal.pgen.1006550>
- Lunt DH (2008) Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins. *BMC Evol Biol* 8:194
- Lynch M (2007) *The origins of genome architecture*. Sinauer Associates Inc., Sunderland
- Lynch M, Bürger R, Butcher D, Gabriel W (1993) The mutational meltdown in asexual populations. *J Hered* 84:339–344
- Lynn DH (2008) *The ciliated protozoa: characterization, classification, and guide to the literature*, 3rd edn. Springer, Dordrecht
- Maciver SK (2016) Asexual amoebae escape Muller's ratchet through polyploidy. *Trends Parasitol* 32:855–862. <https://doi.org/10.1016/j.pt.2016.08.006>
- Maheshwari R (2005) Nuclear behavior in fungal hyphae. *FEMS Microbiol Lett* 249:7–14. <https://doi.org/10.1016/j.femsle.2005.06.031>

- Malik S-B, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM (2008) An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. PLoS ONE 3:e2879
- Marais G (2003) Biased gene conversion: implications for genome and sex evolution. Trends Genet 19:330–338
- Maraun M, Norton RA, Ehnes RB, Scheu S, Erdmann G (2012) Positive correlation between density and parthenogenetic reproduction in oribatid mites (Acari) supports the structured resource theory of sexual reproduction. Evol Ecol Res 14:311–323
- Mark Welch DB, Meselson M (2000) Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. Science 288:1211–1215
- Mark Welch DB, Mark Welch JL, Meselson M (2008) Evidence for degenerate tetraploidy in bdelloid rotifers. Proc Natl Acad Sci USA 105:5145–5149
- Maron JL, Johnson MTJ, Hastings AP, Agrawal AA (2018) Fitness consequences of occasional outcrossing in a functionally asexual plant (*Oenothera biennis*). Ecology 99(2):464–473. <https://doi.org/10.1002/ecs.2099>
- Martinez G, Panda K, Kohler C, Slotkin RK (2016) Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. Nat Plants 2(4). <https://doi.org/10.1038/nplants.2016.30>
- Matzke MA, Kanno T, Matzke AJM (2015) RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. In: Merchant SS (ed) Annual review of plant biology, vol 66, pp 243–267
- Maynard Smith J (1978) The evolution of sex. Cambridge University Press, Cambridge
- Michael TP (2014) Plant genome size variation: bloating and purging DNA. Brief Funct Genom 13(4):308–317
- Mirzaghaderi G, Hörandl E (2016) The evolution of meiotic sex and its alternatives. Proc B-Biol Sci 283(1838). <https://doi.org/10.1098/rspb.2016.1221>
- Mogie M (1992) The evolution of asexual reproduction in plants. Chapman and Hall, London
- Morgens DW, Cavalcanti ARO (2015) Amitotic chromosome loss predicts distinct patterns of senescence and non-senescence in ciliates. Protist 166:224–233. <https://doi.org/10.1016/j.protis.2015.03.002>
- Morrow CA, Fraser JA (2009) Sexual reproduction and dimorphism in the pathogenic basidiomycetes. FEMS Yeast Res 9:161–177. <https://doi.org/10.1111/j.1567-1364.2008.00475.x>
- Muller HJ (1932) Some genetic aspects of sex. Am Nat 66:118–138
- Muller HJ (1964) The relation of recombination to mutational advance. Mutation research 106:2–9
- Neiman M, Sharbel TF, Schwander T (2014) Genetic causes of transitions from sexual reproduction to asexuality in plants and animals. J Evol Biol 27:1346–1359
- Neiman M, Lively CM, Meirmans S (2017) Why sex? A pluralist approach revisited. Trends Ecol Evol 32(8):589–600. <https://doi.org/10.1016/j.tree.2017.05.004>
- Neiman M, Meirmans PG, Schwander T, Meirmans S (2018) Sex in the wild: how and why field-based studies contribute to solving the problem of sex. Evolution 72(6):1194–1203. <https://doi.org/10.1111/evo.13485>
- Ni M, Feretzaki M, Li W, Floyd-Averette A, Mieczkowski P, Dietrich FS, Heitman J (2013) Unisexual and heterosexual meiotic reproduction generate aneuploidy and phenotypic diversity de novo in the yeast *Cryptococcus neoformans*. PLoS Biol 11(9):e1001653. <https://doi.org/10.1371/journal.pbio.1001653>
- Normark BB (1999) Evolution in a putatively ancient asexual aphid lineage: recombination and rapid karyotype change. Evolution [Internet] 53:1458. Available from: <https://doi.org/10.2307/2640892>
- Normark BB, Johnson NA (2011) Niche explosion. Genetica 139:551–564
- Normark BB, Judson OP, Moran NA (2003) Genomic signatures of ancient asexual lineages. Biol J Linn Soc Lond 79:69–84



- Nowell RW, Almeida P, Wilson CG, Smith TP, Fontaneto D, Crisp A et al (2018) Comparative genomics of bdelloid rotifers: insights from desiccating and nondesiccating species. *PLOS Biol* [Internet] 16:e2004830. Available from: <https://doi.org/10.1371/journal.pbio.2004830>
- Oberwinkler F (2017) Yeasts in pucciniomycotina. *Mycol Prog* 16:831–856. <https://doi.org/10.1007/s11557-017-1327-8>
- O’Gorman CM, Fuller HT, Dyer PS (2009) Discovery of a sexual cycle in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Nature* 457:471–475
- Ollivier M, Gabaldón T, Poulain J, Gavory F, Leterme N, Gauthier J-P, Legeai F, Tagu D, Simon JC, Rispe C (2012) Comparison of gene repertoires and patterns of evolutionary rates in eight aphid species that differ by reproductive mode. *Genome Biol Evol* 4:155–167
- Ong-Abdullah M, Ordway JM, Jiang N, Ooi SE, Kok SY, Sarpan N et al (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* 525(7570):533. <https://doi.org/10.1038/nature15365>
- Otto SP (2009) The evolutionary enigma of sex. *Am Nat* 174:S1–S14. <https://doi.org/10.1086/599084>
- Ozias-Akins P, Conner JA (2019) Clonal reproduction through seeds in sight for crops. *Trends Genet* (in press). <https://doi.org/10.1016/j.tig.2019.12.006>
- Ozias-Akins P, van Dijk PJ (2007) Mendelian genetics of apomixis in plants. *Ann Rev Genet* 41:509–537. <https://doi.org/10.1146/annurev.genet.40.110405.09051>
- Parfrey LW, Lahr DJG, Katz LA (2008) The dynamic nature of eukaryotic genomes. *Mol Biol Evol* 25:787–794. <https://doi.org/10.1093/molbev/msn032>
- Parker DJ, Bast J, Jalvingh K, Dumas Z, Robinson-Rechavi M, Schwander T (2019) Repeated evolution of asexuality involves convergent gene expression changes. *Mol Biol Evol* 36:350–364
- Pellino M, Hojsgaard D, Schmutzer T, Scholz U, Hörndl E, Vogel H et al (2013) Asexual genome evolution in the apomictic *Ranunculus auricomus* complex: examining the effects of hybridization and mutation accumulation. *Molec Ecol* 22(23):5908–5921. <https://doi.org/10.1111/mec.12533>
- Peraza-Reyes L, Malagnac F (2016) Sexual development in fungi. In: Wendland J (ed) *The Mycota* I, 3rd edn. Springer, Berlin
- Pöggeler S (2002) Genomic evidence for mating abilities in the asexual pathogen *Aspergillus fumigatus*. *Curr Genet* 42(3):153–160
- Pöggeler S, Nowrousian M, Teichert I, Beier A, Kück U (2018) Fruiting body development in ascomycetes. In: Anke T, Schöffler A (eds) *The Mycota XV, physiology and genetics*, 2nd edn. Springer, Berlin
- Raikov IB (1982) *The protozoan nucleus: morphology and evolution*. Springer, Wien
- Ramesh MA, Malik S-B, Logsdon JM (2005) A phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Curr Biol* 15(2):185–191
- Raudaskoski M, Kothe E (2010) Basidiomycete mating type genes and pheromone signaling. *Eukaryot Cell* 9:847–859. <https://doi.org/10.1128/EC.00319-09>
- Rice WR, Friberg U (2009) A graphical approach to lineage selection between clonals and sexuals. In: Schön I, Martens K, Dijk P (eds) *Lost sex: the evolutionary biology of parthenogenesis*. Springer, Dordrecht, pp 75–97
- Richards CL, Alonso C, Becker C, Bossdorf O, Bucher E, Colome-Tatche M et al (2017) Ecological plant epigenetics: evidence from model and non-model species, and the way forward. *Ecol Lett* 20(12):1576–1590. <https://doi.org/10.1111/ele.12858>
- Ropars J, Toro KS, Noel J, Pelin A, Charron P, Farinelli L et al (2016) Evidence for the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi. *Nat Microbiol* 1(6):16033. <https://doi.org/10.1038/nmicrobiol.2016.33>
- Ross L, Hardy NB, Okusu A, Normark BB (2013) Large population size predicts the distribution of asexuality in scale insects. *Evolution* 67:196–206
- Sampaio JP (2004) Diversity, phylogeny and classification of basidiomycetous yeasts. In: Agerer R, Piepenbring M, Blanz P (eds) *Frontiers in basidiomycete mycology*. IHW, Eching, pp 49–80

- Saunders CW, Scheynius A, Heitman J (2012) *Malassezia* fungi are specialized to live on skin and associated with dandruff, eczema, and other skin diseases. *PLoS Pathog* 8(6):e1002701. <https://doi.org/10.1371/journal.ppat.1002701>
- Schemske DW, Lande R (1985) The evolution of self-fertilization and inbreeding depression in plants. II. Empirical observations. *Evolution* 39:41–52
- Scheu S, Drossel B (2007) Sexual reproduction prevails in a world of structured resources in short supply. *Proc R Soc B-Biol Sci* 274:1225–1231
- Schmidt A, Schmid MW, Klostermeier UC, Qi WH, Guthori D, Sailer C et al (2014) Apomictic and sexual germline development differ with respect to cell cycle, transcriptional, hormonal and epigenetic regulation. *PLoS Genet* 10(7):21. <https://doi.org/10.1371/journal.pgen.1004476>
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Schön I, Martens K (2003) No slave to sex. *Proc B-Biol Sci* 270:827–833
- Schön I, Martens K, van Dijk P (2009) *Lost sex: the evolutionary biology of parthenogenesis*. Springer, Berlin
- Schurko AM, Logsdon JM (2008) Using a meiosis detection toolkit to investigate ancient asexual “scandals” and the evolution of sex. *BioEssays* 30(6):579–589. <https://doi.org/10.1002/bies.20764>
- Schurko AM, Neiman M, Logsdon JM (2009) Signs of sex: what we know and how we know it. *Trends Ecol Evol* 24:208–217
- Schwander T (2016) Evolution: the end of an ancient asexual scandal. *Curr Biol* 26:R233–R235
- Schwander T, Crespi BJ (2009a) Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. *Molec Ecol* 18(1):28–42. <https://doi.org/10.1111/j.1365-294X.2008.03992.x>
- Schwander T, Crespi BJ (2009b) Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. *Mol Ecol* 18:28–42
- Schwander T, Henry L, Crespi BJ (2011) Molecular evidence for ancient asexuality in timema stick insects. *Curr Biol* 21:1129–1134
- Sharbel TF, Voigt M-L, Corral JM, Galla G, Kumlehn J, Klukas C et al (2010) Apomictic and sexual ovules of *boecheera* display heterochronic global gene expression patterns. *Plant Cell* 22(3):655–671. <https://doi.org/10.1105/tpc.109.072223>
- Sharp NP, Otto SP (2016) Evolution of sex: Using experimental genomics to select among competing theories. *BioEssays* 38:751–757
- Shiu PKT, Raju NB, Zickler D, Metzberg RL (2001) Meiotic silencing by unpaired DNA. *Cell* 107(7):905–916
- Signorovitch A, Hur J, Gladyshev E, Meselson M (2015) Allele sharing and evidence for sexuality in a mitochondrial clade of bdelloid rotifers. *Genetics* 200:1–10
- Simon JC, Delmotte F, Rispe C, Crease T (2003) Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals. *Biol J Linn Soc* 79:151–163
- Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8:272–285
- Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijo JA et al (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in Pollen. *Cell* 136(3):461–472. <https://doi.org/10.1016/j.cell.2008.12.038>
- Song Y, Drossel B, Scheu S (2011) Tangled bank dismissed too early. *Oikos* 120:1601–1607
- Spatafora JW, Aime MC, Grigoriev IV, Martin F, Stajich JE, Blackwell M (2017) The fungal tree of life: from molecular systematics to genome-scale phylogenies. *Microbiol Spectr* 5:FUNK-0053-2016
- Speijer D (2016) What can we infer about the origin of sex in early eukaryotes? *Philos Trans Roy Soc B-Biol Sci* 371(1706). <https://doi.org/10.1098/rstb.2015.0530>

- Speijer D, Lukes J, Elias M (2015) Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc Natl Acad Sci USA* 112(29):8827–8834. <https://doi.org/10.1073/pnas.1501725112>
- Stajich JE (2017) Fungal genomes and insights into the evolution of the kingdom. *Microbiol Spectr* 5. <https://doi.org/10.1128/microbiolspec.FUNK-0055-2016>
- Suomalainen E, Saura A, Lokki J (1987) *Cytology and evolution in parthenogenesis*. CRC Press
- Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH (2016) Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. *Genome Biol Evol* 8:2964–2978
- Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R et al (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci USA* 110(50):20117–20122. <https://doi.org/10.1073/pnas.1313452110>
- Tucker AE, Ackerman MS, Eads BD, Xu S, Lynch M (2013) Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. *Proc Natl Acad Sci USA* 110:15740–15745
- Ulum FB, Costa Castro C, Hörandl E (2020) Ploidy-dependent effects of light stress on the mode of reproduction in the *Ranunculus auricomus* complex (*Ranunculaceae*). *Front Plant Sci* 11:104. <https://doi.org/10.3389/fpls.2020.00104>
- Vakhrusheva OA, Mnatsakanova EA, Galimov YR, Neretina TV, Gerasimov ES, Ozerova SG et al (2018) Recombination in a natural population of the bdelloid rotifer *Adineta vaga*. *bioRxiv* [Internet]:489393. Available from: <https://www.biorxiv.org/content/early/2018/12/17/489393>
- van der Kooij CJ, Matthey-Doret C, Schwander T (2017) Evolution and comparative ecology of parthenogenesis in haplodiploid arthropods. *Evol Lett* 1:304–316
- Verhoeven KJF, Preite V (2014) Epigenetic variation in asexually reproducing organisms. *Evolution* 68(3):644–655. <https://doi.org/10.1111/evo.12320>
- Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A (2015) Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol* 23(10):598–605. <https://doi.org/10.1016/j.tim.2015.07.006>
- Wang QM, Begerow D, Groenewald M, Liu XZ, Theelen B, Bai FY, Boekhout T (2015) Multigene phylogeny and taxonomic revision of yeasts and related fungi in the *Ustilaginomycotina*. *Stud Mycol* 81:55–83. <https://doi.org/10.1016/j.simyco.2015.10.004>
- Wang X, Xu YT, Zhang SQ, Cao L, Huang Y, Cheng JF et al (2017) Genomic analyses of primitive, wild and cultivated *Citrus* provide insights into asexual reproduction. *Nature Genet* 49(5):765. <https://doi.org/10.1038/ng.3839>
- Warren WC, García-Pérez R, Xu S, Lampert KP, Chalopin D, Stöck M, Loewe L, Lu Y, Kuderna L, Minx P, Montague MJ, Tomlinson C, Hillier LW, Murphy DN, Wang J, Wang Z, Garcia CM, Thomas GCW, Volff J-N, Farias F, Aken B, Walter RB, Pruitt KD, Marques-Bonet T, Hahn MW, Kneitz S, Lynch M, Schartl M (2018) Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nat Ecol Evol* 2:669–679
- Watts PC, Lundholm N, Ribeiro S, Ellegaard M (2013) A century-long genetic record reveals that protist effective population sizes are comparable to those of macroscopic species. *Biol Lett* 9:20130849. <https://doi.org/10.1098/rsbl.2013.0849>
- Weismann A (1904) *The evolution theory*. Edward Arnold, London
- Weisse T, Anderson R, Arndt H, Calbet A, Hansen PJ, Montagnes DJS (2016) Functional ecology of aquatic phagotrophic protists—concepts, limitations, and perspectives. *Eur J Protistol* 55:50–74. <https://doi.org/10.1016/j.ejop.2016.03.003>
- Wendte JM, Zhang YW, Ji LX, Shi XL, Hazarika RR, Shahryary Y et al (2019) Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. *Elife* 8. <https://doi.org/10.7554/eLife.47891>
- West SA, Lively CM, Read AF (1999) A pluralist approach to sex and recombination. *J Evol Biol* 12:1003–1012
- White MJD (1977) *Animal cytology and evolution*. CUP Archive
- White EP, Ernest SKM, Kerkhoff AJ, Enquist BJ (2007) Relationships between body size and abundance in ecology. *Trends Ecol Evol* 22:323–330

- Williams GC (1975) Sex and evolution. Princeton University Press, Princeton
- Wilson CG, Nowell RW, Barraclough TG (2018) Cross-contamination explains “inter and intraspecific horizontal genetic transfers” between asexual bdelloid rotifers. *Curr Biol* 28:2436-2444.e14
- Wright S, Finnegan D (2001) Genome evolution: sex and the transposable element. *Curr Biol* 11(8):R296–R299. [https://doi.org/10.1016/s0960-9822\(01\)00168-3](https://doi.org/10.1016/s0960-9822(01)00168-3)
- Zhang H, West JA, Zufall RA, Azevedo RBR (2019) Amitosis confers benefits of sex in the absence of sex to *Tetrahymena*. *bioRxiv* 794735
- Zhao Y, Wang Y, Upadhyay S, Xue C, Lin X (2020) Activation of meiotic genes mediates ploidy reduction during cryptococcal infection. *Curr Biol* 30:1387-1396.e5. <https://doi.org/10.1016/j.cub.2020.01.081>
- Zickler D, Espagne E (2016) *Sordaria*, a model system to uncover links between meiotic pairing and recombination. *Semin Cell Dev Biol* 54:149–157

# Chapter 8

## On the Origin of Life and Evolution of Living Systems from a World of Biological Membranes



Aditya Mittal, Suneyna Bansal, and Anandkumar Madhavjibhai Changani

**Abstract** The central dogma, i.e. DNA to RNA to protein, is central to biology. Biological membranes are also central to biology. However, discussions on origin of life and evolution of living systems rely primarily on nucleic acids and proteins. A contextual appreciation of biological membranes in evolutionary biology has not yet emerged. One of the primary reasons for this is the replication mechanism offered via DNA and its subsequent transcription and translation. In this work, we explore the possibility of a replication mechanism in nucleic acid-free and protein-free self-assembled systems based on the law of mass action. While exploring the role of water, both as a chemical reactant and as a solvent, we present the view of a “micellar world” towards understanding origin of life and subsequent evolution through a thought experiment that we call Life of Micellar Systems (LoMS). We hope that the ideas presented will stimulate discussions on biological membranes towards building completely new perspectives not only in evolutionary biology but also in synthetic biology.

---

A. Mittal (✉) · S. Bansal · A. M. Changani  
Kusuma School of Biological Sciences, Indian Institute of Technology Delhi (IIT Delhi), Hauz Khas, New Delhi 110016, India  
e-mail: [amittal@bioschool.iitd.ac.in](mailto:amittal@bioschool.iitd.ac.in)

A. Mittal  
Kusuma School of Biological Sciences and Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi, Hauz Khas, New Delhi 110016, India

S. Bansal  
NIIT Technologies Limited, Tech Zone IT City, Greater Noida, Uttar Pradesh, India

A. M. Changani  
Growth Source Financial Technologies, Mumbai, Maharashtra, India

## 8.1 Biological Membranes and Evolutionary Biology

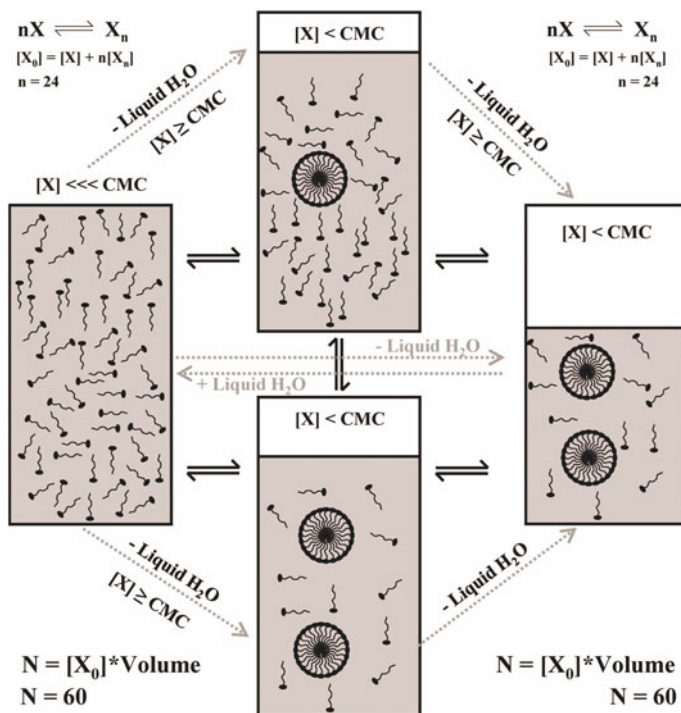
Biological membranes are central to biology. Starting from the fluid mosaic model (Singer and Nicolson 1972) to the modern description of biological membranes as two-dimensional fluids with domains/rafts of varying sizes (Brown and London 1998), the overall view of biological membranes has evolved over the years. From a perspective of simply serving as semi-permeable boundaries for whole cells and their intracellular compartments, it is now well-appreciated that there are several key roles played by lipid constituents in biological membranes—for example, molecular recognition, signalling, trafficking and subcellular (re)organization have been discovered in the last couple of decades (see introduction in Bansal and Mittal 2013). However, when it comes to discussions on origin of life and evolution of living systems, biological membranes almost do not find a comprehensive mention. Such discussions are essentially dominated by (self) catalyzing RNA and/or proteins (Deamer 2019). This is because the hallmark of living systems is considered to be the intracellular operation of the central dogma (DNA to RNA to protein)—biological membranes do not feature in the central dogma. Thus, in spite of availability of abundant literature on the informatics of nucleic acids and proteins, “lipidomics” is mostly addressed in literature in terms of enzymes that carry out (bio) chemical transformations resulting in formation of lipids. In one of the first analyses inspired by computational genomics and proteomics, a model for the origin of eukaryotic cells was developed based on membrane lipidomics on three domains of life—a unique lipid class diagram was developed for classifying thousands of lipids that constitute biological membranes in all the three domains of life (Bansal and Mittal 2015). Subsequently, analyses of thickness of subcellular biological membranes allowed development of some interesting novel insights into origins and evolution of cellular and multi-cellular life forms (Singh and Mittal 2016; Mittal and Singh 2018). However, evolutionary biology is still far away from giving considerations to biological membranes in comparison with nucleic acids and proteins.

From invention of optical microscopy up to mid-twentieth century, erythrocytes (red blood cells—RBCs), in addition to having their own significance in clinical and basic biology, have served as model systems for insights into behaviour of biological membranes (Singh et al. 2019). Subsequently, liposomes have served as a promising model system for understanding biological membranes for the last half-a-century. However, there is only one report till date that experimentally derives a quantitative formalism called the “Critical Compartmentalization Concentration” for stoichiometric assembly of amphipathic molecules into aggregating assemblies of a given size that serve as compartments (Mittal and Grover 2010). In contrast, assembly of micelles from amphipathic molecules is extremely well covered in the literature, largely owing to the pioneering work of Tanford (1973, 1978) on the hydrophobic effect. Micellar aggregation also serves as a model system towards gaining insights into formation and behaviour of biological membranes. Thus, in this chapter, we propose a completely new model for the origin of life and evolution of living systems based entirely on micellar systems. By carrying out a thought experiment called

“Life of Micellar Systems” (LoMS), we propose a “micellar world” for origin of life and evolution of living systems rather than a primarily “protein” and/or “RNA” world. We are hopeful that the ideas presented will stimulate discussions on biological membranes from a completely new perspective contributing towards not only evolutionary biology but also synthetic biology.

**The Life of Micellar Systems (LoMS) “thought” experiment** Here we discuss the origin of life and evolution of living systems from a “Micellar World”. Consider a system shown in the leftmost panel of Fig. 8.1.

Micelle forming amphiphiles (denoted by “X”), with a hydrophilic head group and a hydrophobic chain, are suspended in a given amount of water—water serves primarily as the solvent. As per the law of mass action, aggregation of X resulting in a micelle is shown by the equation given at the top of Fig. 8.1—a key assumption, for simplification, here is that there are only two possible thermodynamic states in which the molecular species “X” can exist: either a monomer or a micelle (no other aggregate is considered to be either thermodynamically favourable or have any kinetic lifetime to be of any consequence). Further, assume that the aggregation number of



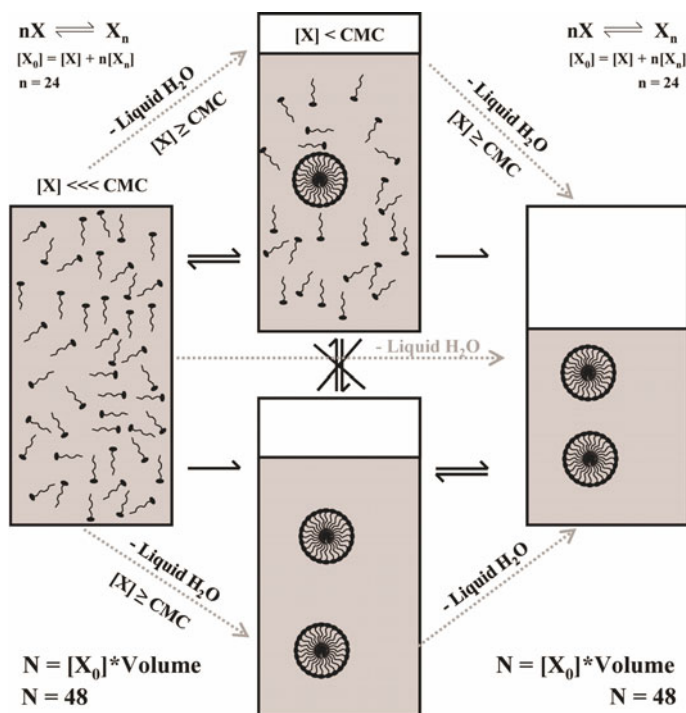
**Fig. 8.1 Birth of a micelle**—aggregation of amphiphiles in water and apparent “replication” of aggregates by law of mass action. The mass balance on X is given by  $N = [X_0] * Volume$ , where  $[X_0] = [X] + n[X_n]$

$X$ , denoted by “ $n$ ”, is 24 (i.e. formation of a micelle requires exactly 24 molecules of  $X$ ). Total number of  $X$  molecules in the system, denoted by “ $N$ ”, is 60 and remains constant. In the leftmost panel of Fig. 8.1, the concentration of monomeric  $X$ , given by  $[X]$ , is much lower than the critical micellar concentration (CMC), i.e.  $[X] \lll \text{CMC}$ . Now, let us remove some water from the system—the top middle panel in Fig. 8.1 shows the formation of single micelle transiently due to the concentration of  $X$  reaching the threshold of CMC due to the removal of water—here it is important to note that removal of water as a solvent is not to be confused with “dehydration” of the micelles; dehydration at a molecular level results in conformational alterations and changes in aggregation patterns. In fact, it is well known that bulk water properties (water serving as a solvent) are significantly different from the properties of water that form the hydration layers in the self-assembled structures.

A little more removal of water, shown by the bottom middle panel in Fig. 8.1, shows the formation of two micelles transiently—in both these systems, there is an equilibrium between  $X$  (monomers) and  $Xn$  (micelles), depending on the amount of water in the system. Note that since  $n = 24$  and  $N = 60$ , at most two micelles are formed in this system. Now, further the removal of water from the system, as shown by the rightmost panel in Fig. 8.1, will not affect the equilibrium between the two micelles and the monomers. Thus, we have observed the spontaneous “birth of a micelle” in a system simply by the manifestation of the law of mass action in a system where the quantity of water as a solvent is varied. Note that the transitions shown by only removal or addition of water, and maintenance of  $N = 60$  in the system, are all reversible due to the law of mass action with corresponding addition or removal of water. However, in very dilute conditions, i.e. if  $[X] \lll \text{CMC}$ , no micelle will ever form (as in leftmost panel). Similarly, at high concentration of  $X$  (but  $N$  still = 60) due to the removal of sufficient water, two micelles will always be at equilibrium with 12 monomeric  $X$  (since  $n = 24$ ). Clearly, this not only provides a direct analogy to spontaneous appearance of a single self-assembled system (“birth of a micelle”) further leading to appearance of “replication” resulting in appearance of two similar (in this case identical) self-assembled systems simply by controlling water volume. Now consider a system shown in the leftmost panel of Fig. 8.2.

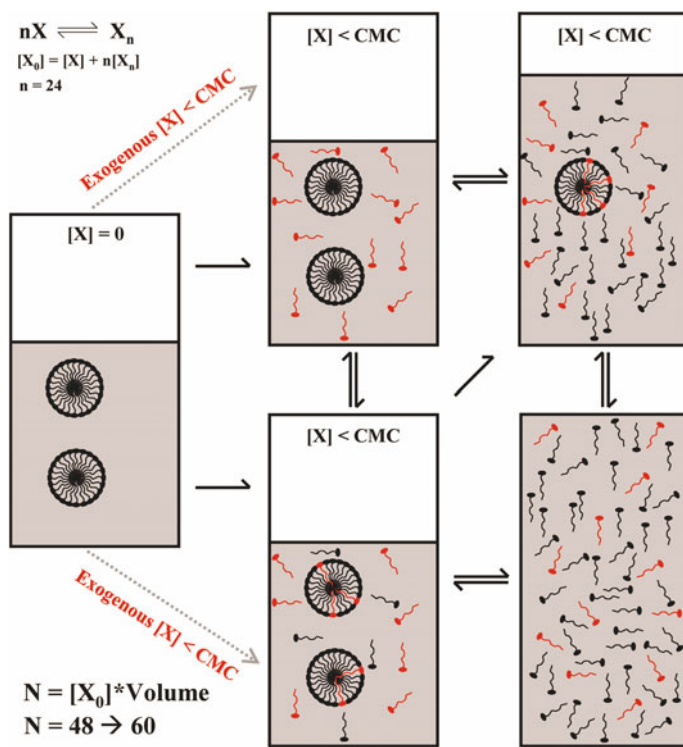
All of the system characteristics are exactly the same as those in Fig. 8.1, except that  $N = 48$  in this system (instead of  $N = 60$  as in Fig. 8.1). Therefore, here the total number of  $X$  molecules in the system is 48 and remains constant. Again, the leftmost panel of Fig. 8.2 shows that the concentration of monomeric  $X$ , given by  $[X]$ , is much lower than the critical micellar concentration (CMC), i.e.  $[X] \lll \text{CMC}$ . Now, removal of some water from the system, as shown in the top middle panel of Fig. 8.2, results in formation of single micelle transiently due to the concentration of  $X$  reaching the CMC threshold. A little more removal of water, shown by the bottom middle panel in Fig. 8.2 shows the formation of two micelles; however, micelle formation is not transient in both the systems this time. This is because the number of monomers is an exact multiple of the aggregation number of the monomer; hence, once the energetically favourable micellar configuration is reached for all monomers (i.e. micelles formed do not disaggregate—this results in disappearance of monomers). Of course, the law of mass action may now manifest itself via exchange of amphiphilic





**Fig. 8.2 Life of a micelle**—aggregation of amphiphiles in water, the illusion of replication and lifespan of aggregates in a “stagnant” system. The mass balance on  $X$  is given by  $N = [X_0] * \text{Volume}$ , where  $[X_0] = [X] + n[X_n]$

molecules between the two micelles in the bottom middle panel; however, there will be no remaining monomeric species. Thus, there is a clear distinction in this case from that shown in Fig. 8.1—while there is an equilibrium between  $X$  and  $X_n$  in the top middle panel, the bottom middle panel represents a stable and irreversible micellar system with complete absence of monomers. Further, the removal of water (shown in the rightmost panel) or even further addition of water to the system in the bottom middle panel will not affect the system since the micellar aggregates are energetically stable. These observations can be specifically noted by contrasting the reversibility versus irreversibility of the systems, represented by reaction directions, in Figs. 8.1 and 8.2. Thus, here we have observed not only the spontaneous “birth and replication of micelles” but also “lifespan of micelles”. This lifespan of the micelles will continue until (a) there is an energetic epoch in the system to destabilize one or both of the micelles, or (b) there is another molecular species that can interact with  $X$  is introduced in the system that can have an “interaction constant” (e.g. binding affinity) to extract  $X$  out of either or both micelles, or (c) there is an introduction of more  $X$  molecules in order to re-invoke the law of mass action-driven equilibrium between monomers and micelles. Until such perturbations are introduced in the system, the



**Fig. 8.3 Death of micelles**—Disaggregation of stable micelles due to perturbations. The mass balance on  $X$  is given by  $N = [X_0] * \text{Volume}$ , where  $[X_0] = [X] + n[X_n]$

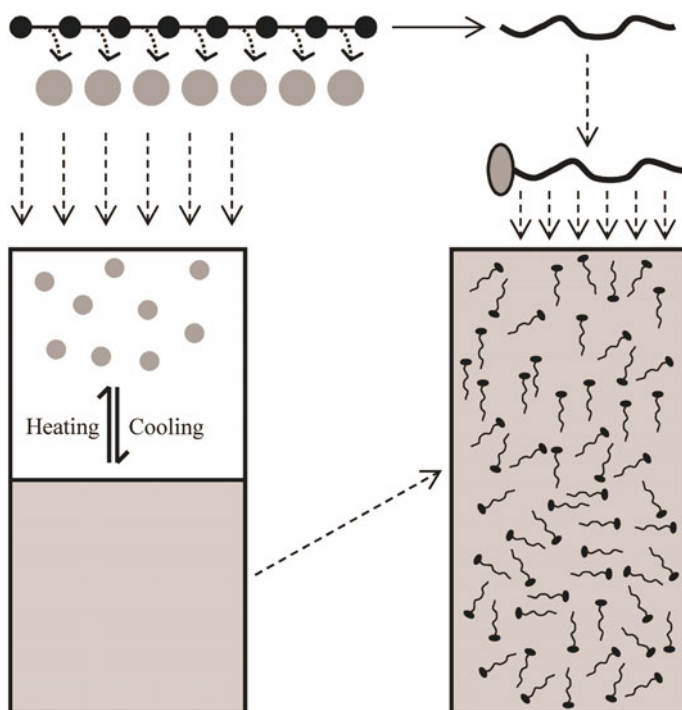
two micelles will apparently “live happily forever after”. Now, let us introduce one of the perturbations discussed above—the simplest one being introduction of more  $X$  monomers in the system. Figure 8.3 shows addition of more monomeric  $X$  molecules in the “living system” shown in the leftmost panel. The “living system” has  $N = 48$  (same as in Fig. 8.2), and 12 more monomers are added to bring the total molecules of  $X$  to  $N = 60$  (same as in Fig. 8.1). For clarity, the new  $X$  monomers are shown in red, compared to the original black. Without any change in the amount of water, the system will be driven towards equilibrium between monomers and micelles driven by the law of mass action. Now, after addition of the new monomers, further addition of water will result in dilution of the system and at some point when  $[X] \ll \text{CMC}$ , the system will resemble the original starting system in Fig. 8.1 as shown here in the right bottom panel.

Through these straightforward thought experiments, we show that the apparent cycles of birth–replication–life–death are simply manifestations of the law of mass action with water serving as a solvent for different molecular species. Thus, we propose that not only did life actually originate in this manner, but it continues

to evolve similarly with different molecular species resulting in different kinds of aggregates of different sizes and quantities driven by the law of mass action.

Having proposed a theory for origin of life and evolution of living systems based on the law of mass action, it becomes important for us to address the question of raw materials for the system. As mentioned earlier, while substantial literature focuses on nucleic acids or proteins or both in form of self catalyzing systems for origin of life, we propose amphiphilic assembly (along with, or later with the addition of, a plethora of chemical reactions) in water as a solvent for origins of life. We propose that abundance of carbon on Earth coupled with high-temperature conditions allowed catenation reactions resulting in chains of carbon being formed (black circles in Fig. 8.4). Formation of carbon chains by condensation reactions resulted in water molecules (grey circles, Fig. 8.4)—this water formed was still in vapour form due to the conditions that were resulting in formation of carbon chains.

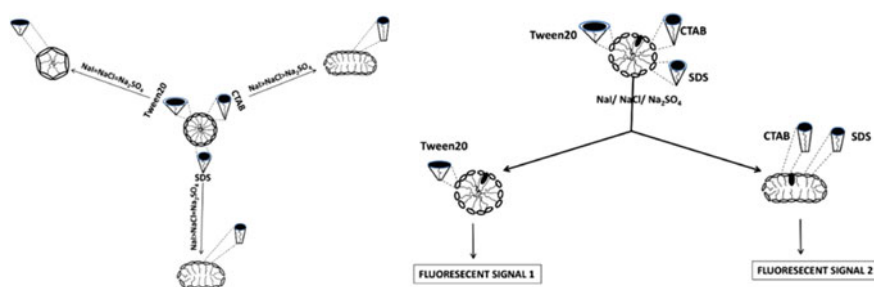
Thus, before life originated, Earth's atmosphere was rich in gaseous form of carbon-chained molecules along with water vapour along with other molecules in gas phase. With high water vapour content, and heavier carbon-based molecules (including many amphiphiles with carbon chains) having lower volatility, heating and cooling cycles resulted in the formation of liquid pools of water with populations



**Fig. 8.4 Emergence of conditions for origin of life and evolution**—formation of water as a solvent and reaction medium for self-assembling reactions

of amphiphilic molecules. Of course, other chemically reacting systems also must have formed. However, the predominant occurrence of amphiphiles assembling and disassembling in water resulted in the first aggregating systems—these aggregations driven by the energetics of the hydrophobic effect also resulted in what is currently apparent as replication. Simultaneous addition and removal of different molecular species resulted in emergence and sustenance of newer aggregate forms. Here, it is interesting to note that formation of primary polymer chains in living cells, i.e. of nucleic acids and proteins, actually consume water—in test-tube chemistry, polymerization reactions actually result in release of water. Now in the actual intracellular conditions, polymerization reactions resulting in water as a product to form polymer chains cannot take place in aqueous milieu. The presence of water as a solvent at 55 M (concentration of liquid water) would actually drive the reaction backward by law of mass action thereby inhibiting polymerization and favouring the backward reaction due to excess product (water). Operation of the central dogma in living cells actually consumes water—we calculate the amount of this water consumed in Appendix 1. From an evolutionary perspective, this is a very important discovery—once amphiphilic assembly resulted in apparently replicating structures in excess water, more chemical reactions with different molecular species but still driven by the law of mass action took over towards emergence of living cells in their current avatar. Many of these cellular chemical reactions, and certainly the key ones for living system, involve water as a chemically reacting molecular species and not just as a solvent. Finally, having appreciated the role of water as a chemically re(active) molecular species, and not just as a solvent, we also show how different and reproducible morphologies of micellar aggregates emerge in different aqueous chemical environments; see Fig. 8.5 and Appendix 2.

Concluding, we propose that origin of life and evolution of living systems are essentially a result of both the kinetics and the thermodynamics of the hydrophobic effect rooted in the law of mass action, with water as a solvent. While this chapter is aimed at a simplistic introduction of a somewhat orthogonal view on the origin of living systems, it is our hope that self-assembly driven primarily by solvent characteristics will emerge as a key driver of hypotheses in understanding living systems.



**Fig. 8.5** Regulation of amphiphilic self-assembly by water structure and molecular shapes (see Appendix 2 for details)

For this, further studies on the role of the solvent species, both as a reactant and as a reaction medium, will shed further light on origin–sustenance–demise cycles of self-replicating structures as models for origins of life and evolution. Specifically, our proposal emphasizes the importance of amphiphiles and biological membrane-like materials resulting in the first “living” systems rather than nucleic acids and/or proteins. In essence, we are proposing a “micellar world” for origin of life and evolution of living systems rather than a “protein and/or RNA world”. One very interesting and appealing aspect of the ideas presented here is that it may not be required to treat understanding of origin of life as a “chicken and egg” problem, if viewed from the perspectives discussed in this work.

**Acknowledgements** SB is grateful to the Council of Scientific and Industrial Research, Government of India, for research fellowship support. AMC is grateful to IIT Delhi for research fellowship support. AM is grateful to Marie-Hélène Rome and Pierre Pontarotti for their patience.

**Author Contributions** SB carried out the experimental work for Appendix 2 and co-wrote Appendix 2 with AM, including preparation of figures. AMC carried out literature review and collected the data required for Appendix 1. AM designed the study, formulated the micellar world hypothesis and Life of Micelle Systems thought experiment and wrote the manuscript.

## Appendix 1: Water Consumption in Intracellular Central Dogma Operations

**Summary** Carbon-based chemical polymerization releases water molecules. In contrast, intracellular biopolymerization actually consumes water. Surprisingly, water requirements for intracellular synthesis of proteins and nucleic acids are not accounted for in literature. In this work, we derive the first quantitative expression for the number of water molecules consumed per molecule of protein produced by transcription and translation in a cell. Extrapolating our findings to multi-cellular organisms, we find that a staggering ~90 million litres of water per day is utilized by the current human population only for synthesizing intracellular proteins for survival without replication/reproduction—and this is an underestimate.

**Introduction** Water as a solvent is assumed to be a prerequisite for life. Not only does water serve as a solvent for chemical reactions inside and outside living cells, but it also serves as a solvent that drives physical (and often non-chemical) self-assembly of amphipathic molecules via the “exclusion by water” principle formally referred to as the hydrophobic effect (Tanford 1973; Silverstein et al. 1998; Southall et al. 2002; Dill et al. 2005; Xu and Dill 2005; Mittal and Grover 2010; Dill and Bromberg 2011). In fact, elucidation of unique properties of water as a solvent for life has been of constant interest (Truskett and Dill 2003; Urbic et al. 2007; Fennell et al. 2010; Mittal and Jayaram 2011; Urbic and Dill 2017; Brini et al. 2017; Urbic and Dill 2018). While appreciating the role of water as a solvent for physical, chemical and physico-chemical interactions governing living systems, it is also pertinent

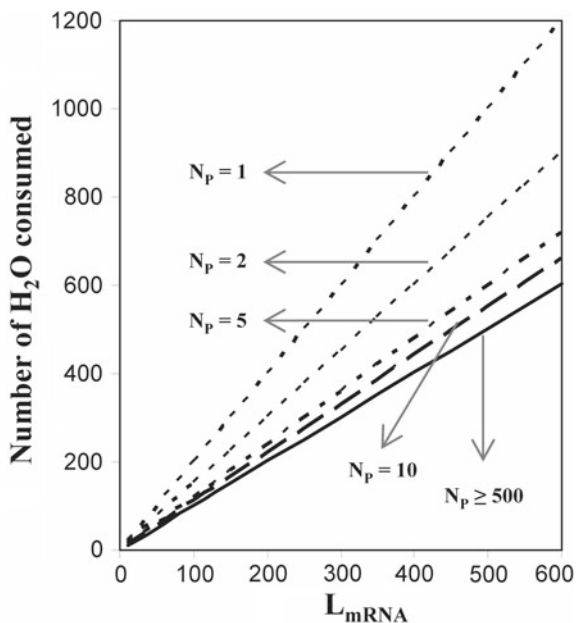
to consider its role as a chemical reactant in living systems. For example, purely chemically driven polymerization reactions involving formation of carbon-based chains involve release of a water molecule per bond formed. Thus, in such systems, the presence of water as a solvent would not allow a forward reaction resulting in release of water in order to form a bond for extending the polymer due to the high concentration (55.5 M) of surrounding water. However, interestingly in living cells, polymerization of nucleotides resulting in the formation of DNA and mRNA molecules as well as synthesis of polypeptide chains (resulting in proteins) actually utilize water as a reactant for bond formation required for polymer elongation. Surprisingly, this water requirement during the creation of biosynthetic machinery in living cells is largely ignored and not accounted for in the literature, including textbooks on life and cell biology. The apparent assumption of the constitutively present biosynthetic machinery from the perspective of requirements of water as a reactant is in complete contrast to substantially increasing literature on energy requirements for assembling the same biosynthetic machinery. In fact, energetic considerations in biological systems (e.g. living cells) continue to be addressed regularly with respect to cell size and growth (Dill et al. 2011; Naresh et al. 2012; Maitra and Dill 2015; Szenk et al 2017; Wagoner and Dill 2019) along with assembly of, and functional constraints on, proteomes and cell organelles (Ghosh and Dill 2010; Naresh et al. 2010; Ghosh et al. 2016; Santra et al. 2017; Agozzino and Dill 2018). Therefore, in this work, we explicitly address the requirement of water in creation of the biologically synthesized machinery (i.e. proteins). We derive the first-ever formalism for calculating water consumed during the processes of transcription and translation resulting in synthesis of a protein molecule. Our results compel explicit inclusion of water consumed during operation of the central dogma for creation of biosynthetic machinery during considerations of metabolic perspectives on living systems.

## ***Results and Discussion***

**Accounting for water consumed during transcription and translation** First, let us consider transcription. Formation of a phosphodiester bond to extend an mRNA molecule requires one water molecule to hydrolyze the released pyrophosphate (Heinonem 2001; Nelson and Cox 2008). Therefore, addition of each nucleotide to form an mRNA molecule consumes one molecule of water (Heinonem 2001; Nelson and Cox 2008). This implies that synthesis of an mRNA molecule with length = " $L_{\text{mRNA}}$ " in terms of number of nucleotides consumes a total of " $L_{\text{mRNA}} - 1$ " water molecules (since the number of phosphodiester bonds is one less than the total length due to the 5' and 3' termini). Now, each mRNA transcript can be utilized to produce a variable number of protein molecules, generally ranging from 10 to 10,000 (Heinonem 2001; Dill et al. 2011; Lahtvee et al. 2017). Let us assume that the number of protein molecules per mRNA transcript is given by " $N_P$ ". Note that each of these protein molecules will have " $L_{\text{mRNA}}/3$ " residues (since codon size =

3). Therefore, the number of water molecules consumed during transcription for a single protein molecule =  $“(L_{\text{mRNA}} - 1)/N_{\text{P}}”$ . The next step is translation. Beyond the presence of an mRNA transcript, translation requires aminoacyl-tRNA molecules. For producing one aminoacyl-tRNA molecule, the first step is activation of an amino acid molecule—this involves hydrolysis of a pyrophosphate by pyrophosphatase which consumes one water molecule (Heinonem 2001; Nelson and Cox 2008; Yang et al. 2009). The activated amino acid, in the form of amino acid-AMP complex, combines with tRNA resulting in a molecule of aminoacyl-tRNA and release of an AMP (Voet et al. 2013). The released AMP is converted into an ADP (Poos et al. 1999). Therefore, formation of a single aminoacyl-tRNA consumes one molecule of water. To extend the polypeptide chain during translation, i.e. for polypeptide elongation, an aminoacyl-tRNA binds with the GTP-bound elongation factor EF-Tu to form a ternary complex. This ternary complex binds to the ribosomal A-site—during this reaction the GTP is hydrolyzed and one water molecule is consumed (for GTP hydrolysis). In addition, polypeptide elongation also involves hydrolysis of EF-G-GTP complex at the ribosomal P-site for removal the uncharged tRNA. This hydrolysis consumes one water molecule. Therefore during the incorporation of a single amino acid into an elongating polypeptide chain, starting with formation of aminoacyl-tRNA to the release of uncharged tRNA, three water molecules are consumed. The above is applicable to all amino acids incorporated into an newly forming polypeptide chain except for the first and the last. An initiation complex involving mRNA, 30S ribosome and GTP-fMet-tRNA is formed by hydrolysis of a GTP molecule (thereby consuming one water molecule) for starting translation (Voet et al. 2013). Thus, formation of this initiation complex consumes one water molecule for starting translation (instead of consuming one water molecule for creation of aminoacyl-tRNA). Subsequent steps remain the same—therefore, a total of three water molecules are consumed for starting translation with the first residue. For the last amino acid incorporated in the polypeptide chain, ribosomal hydrolysis of peptidyl-tRNA resulting in the released polypeptide chain along with free tRNA consumes one water molecule (instead of hydrolysis of EF-G-GTP complex at the ribosomal P-site for removal the uncharged tRNA). Therefore, three water molecules are consumed for ending elongation with the incorporation of last residue. Summarizing requirements during translation, the total number of water molecules consumed to create a polypeptide chain with total residues =  $“L_{\text{mRNA}}3”$  during translation is  $“3 \times L_{\text{mRNA}}3” = “L_{\text{mRNA}} \cdot 3”$ . In addition, two water molecules get consumed during termination of translation while releasing the ribosomal subunits and mRNA transcript immediately subsequent to (or simultaneously with) release of the polypeptide chain (Voet et al. 2013). Therefore, the total number of water molecules consumed in the translation step of creating a single polypeptide chain (total residues =  $“L_{\text{mRNA}}3”$ ) is  $“L_{\text{mRNA}} + 2”$ . Combining transcription and translation, the total water molecules consumed for synthesizing one protein molecule having  $“L_{\text{mRNA}}/3”$  residues =  $[(L_{\text{mRNA}} - 1)/N_{\text{P}}] + [L_{\text{mRNA}} + 2]$ . Thus,

$$\frac{\text{H}_2\text{O molecules consumed}}{\text{Protein molecule}} = \left(1 + \frac{1}{N_{\text{P}}}\right)L_{\text{mRNA}} + \left(2 - \frac{1}{N_{\text{P}}}\right) \quad (8.1)$$



**Fig. 8.6** Total number of water molecules consumed for synthesizing (transcription + translation) a protein molecule with “ $L_{mRNA}/3$ ” residues as a function of the length of the mRNA transcript (in terms of number of nucleotides) encoding for the protein molecule. “ $N_p$ ” is the number of protein molecules synthesized by translating a single mRNA transcript. The plot is based on Eq. (8.1), see text for details

where “ $L_{mRNA}$ ” is the length of the mRNA molecule for synthesizing a protein molecule with “ $L_{mRNA}/3$ ” residues and “ $N_p$ ” is the number of protein molecules synthesized per mRNA transcript. Figure 8.6, based on Eq. (8.1) above, shows that beyond “ $N_p = 500$ ”, there is negligible change in number of water molecules consumed per molecule of protein synthesized regardless of the size of the protein.

**Calculating water consumed only for transcription and translation by a cell**

Having analytically derived the expression giving water consumption during protein biosynthesis, the next step is to actually apply it to in vivo scenarios. To do so, let us consider *E. coli* cells. The average length of a protein in an *E. coli* cell is 325 residues (Zhang 2000; Dill et al. 2011). Therefore, on an average,  $L_{mRNA}/3 = 325 \geq L_{mRNA} = 975$ . Assuming  $N_p = 50$  (Heinonem 2001; Dill et al. 2011; Lahtvee et al. 2017), the total number of water molecules consumed by an *E. coli* cell for synthesizing one protein molecule =  $[1 + (1/50)] \times 975 + [2 - (1/50)] = 996.48$ .

An *E. coli* cell has ~3 million protein molecules (15), and a replication time of ~10 min (15). Therefore, the total water consumed by a non-replicating *E. coli* cell for only surviving for 10 min by synthesizing its protein molecules =  $3 \times 10^6 \times 996.48 = 2.98944 \times 10^9$  molecules. Now there are several ways of looking at this interesting result. As an example, let us consider an *E. coli* culture of  $10^8$  non-replicating cells



per ml—note that is at least an order of magnitude less than a saturated *E. coli* culture (Sezonov et al. 2007). Ten litres of this culture would contain  $1 \times 10^{12}$  cells, and therefore, the total water consumed per day by this culture for only cell survival by protein synthesis would be  $(24 \times 60/10) \times 2.98944 \times 10^9 \times 1 \times 10^{12} = 4.3047936 \times 10^{23}$  molecules. This equals  $(4.3047936/6.023) \sim 0.715$  mol of water. One mole of water is 18 ml of liquid water. Therefore, a 10 L culture of non-replicating *E. coli* cells would consume  $18 \times 0.715$  ml  $\sim 13$  ml of liquid water per day only for protein synthesis required for survival. While 13 ml of water per day might appear a small amount, considering simply a stationary batch secondary-metabolite industrial production culture in a reactor of 20 kilolitres (i.e. cells are in pseudo-steady state in stationary phase without replication), it translates to consumption of  $[(13 \text{ ml} \times 20,000 \text{ L})/(10 \text{ L})] = 26 \text{ L}$  of water by cells only for protein synthesis required for survival. These 26 L of water are not accounted for anywhere in any literature till date. More importantly, since 26 L is  $\sim 0.13\%$  of 20 kilolitres, it never gets noticed. As another example, let us extrapolate the results from *E. coli* cells to multi-cellular organisms such as humans. If we consider a human to be equivalent to a trillion ( $10^{12}$ ) non-replicating *E. coli* cells, then again, a stagnant human population of  $\sim 7$  billion would be consuming a staggering  $[(13 \text{ ml} \times 7 \times 10^9)] = 91$  million litres of water per day just for protein synthesis required for survival without any growth. Of course, this calculation provides an underestimate; however, it strongly highlights the fact that water consumption during transcription and translation needs to be accounted for during our considerations of cell biology and life in general.

**Present and future importance of water consumption during intracellular biosynthesis** We have derived the first estimates of water consumption during transcription and translation per protein molecule synthesized inside living cells. In modern times, with the availability of whole transcriptome and proteome data especially for different cells growing under different conditions, it would be indeed interesting to calculate the water consumed per “unit functionality” of cells. By knowing the exact number of mRNA transcripts and the number of corresponding protein molecules synthesized specifically during a particular metabolic process of a cell (e.g. glucose metabolism or lipid biosynthesis or even more generalized concept of cell growth), water consumption for creation of biosynthetic machinery required during that specific metabolic process can now be calculated. Our work promises to be a useful tool in planning requirement of water resources for setting up and maintenance of living societies (of both microbial and multi-cellular organisms including humans) on Earth, since water is fast becoming a limiting resource, and even beyond since water requirements will have to be calculated accurately.

## Appendix 2: Regulation of Amphiphilic Self-assembly by Water Structure and Molecular Shapes

**Summary** Structural diversity in lipids maintains the dynamic structure and functions of variably curved membranes in cellular milieu. Distribution of molecular shapes of these lipids drives curvature formation in the membrane structures. In aqueous environment, these amphiphiles are known to form preferred assembled geometries due to the hydrophobic effect. In order to gain insights into the dynamics of individual shapes of amphiphiles accompanying (and possibly leading to) this transition in entire assembled structures, we used tryptophan octyl ester (TOE-probe)-based fluorescence assay on three types of surfactants (anionic, cationic and non-ionic) and alter the structure of water using three different electrolytes. The pKa values, fluorescence and red shift of the TOE incorporated into different micellar structures provide a somewhat direct measure of the role of individual molecular shapes of surfactants resulting in various self-assembled structures in different aqueous environments. Here, we provide evidence that perturbation in water structure alters the individual molecular shape and thus, results in shape transition(s) of micellar structures. However, these changes in molecular shapes are allowed within permissible ranges as dictated by chemical structure of the individual amphiphiles. Experimental findings in this work allow an investigation of not only alteration of molecular shapes of individual surfactant molecules in different aqueous environments, but also the role of these alterations in governing the overall structure of the hydrophobic effect-driven macromolecular assemblies.

**Introduction** Membranes are highly active structures present as plasma membrane and organellar membranes in the cells. Dynamic nature and varied functionality in different membranous compartments are maintained by diversity in their lipid compositions (Van Meer et al. 2008; Andreyev et al. 2010, Bansal and Mittal 2015). In addition to various curvature forming proteins (Zimmerberg and Kozlov 2006; Bansal and Mittal 2013), asymmetric distribution of individual molecular shapes of lipids has been known to induce curvature formation in membranes (McMahon and Gallop 2005). These molecular shapes of amphiphiles have been quantified in terms of “packing or shape parameter” by Israelachvili et al. (1976). There are several reports which support that these shape parameter of membrane lipids dictates the self-assembly of amphiphiles in solvent medium. Experimental evidence showed that change in the membrane composition of RBC with differentially shaped lipid molecules could alter the entire cellular morphology (Christiansson et al. 1985). Membrane remodelling to form differentially curved membranes as required during cellular fission or fusion processes (Chernomordik et al. 1985; Chernomordik and Kozlov 2003), tubules or vesicles formation (Roux et al. 2005; Christian et al. 2009), sorting of lipids and membrane proteins (Mukherjee et al. 1999) and fusion of enveloped virus to host membranes (Chernomordik et al. 1995; Mittal et al. 2002; St Vincent et al. 2010; Zaitseva et al. 2010) have been known to be regulated by the array of various lipids. Few computational studies also reported the coupling between lipid shape and geometry of its assembled structure using bead-based model in molecular

dynamics (MD) simulations (Cooke and Deserno 2006) and shape-based phase separation of lipids using conical linactants segregated at the phase boundary (Schafer and Marrink 2010). Subsequent to development of liposomes as model systems by Bangham (1972), there are very few reports which investigate the mechanism of liposomal assembly (Mittal and Grover 2010). However, micelles have served as a major model system to gain mechanistic insights into the self-assembly process of amphiphiles in aqueous solutions till date. As it is well known that the self-assembly of amphiphiles is driven by the hydrophobic effect in aqueous water (Tanford 1973, 1978), we proposed a testable hypothesis to see the effect of any disturbance created in the structure of liquid water on the molecular shapes of amphiphiles and thereby affecting the whole self-assembly process. Water in liquid form is known to be in disordered state having very short range ordered structure as compared to its crystalline state (Tanford 1973). The most accepted model of water structure was given by Eisenberg and Kauzmann (1969) which assumes that water molecules in liquid-state form four hydrogen bonds to their neighbouring water molecules that result a flexible irregular network. Structure of water is known to be affected by the presence of other molecules and electrolytes. Few of the electrolytes have been reported to cause more ordering in the water structures and classified as chaotropes (or water structure breakers) such as NaI and kosmotropes (or water structure makers) such as NaCl, Na<sub>2</sub>SO<sub>4</sub> (Nickolov and Miller 2005; Barbosa et al. 2010). Hofmeister series illustrates the order of the effect of these ions on the structure of water (Collins 1997; Marcus 2009; Mahler and Ingmar 2012) and has been extensively studied to explain their effect on the conformation and stability of proteins (Baldwin 1996; Collins 2004). Kosmotropes, also termed as salting-out salts, are known to decrease the solubility of proteins and cause protein precipitation. Chaotropes are known to increase the solubility of the proteins by increasing their solvent accessible area and termed as salting in salts (Wetlaufer et al. 1964; Zhang and Cremer 2006). Few reports also studied the effect of these electrolytes on micellar systems. Addition of electrolytes has shown to decrease the critical micellar concentration (CMC) of various surfactants indicating their transition from spherical to rod shaped micellar assemblies (Ray and Nemethy 1971; Ericsson et al. 2004). These shape transition in micelles have been indirectly visualized using environment-sensitive fluorescence assays (Rawat and Chattopadhyay 1999; Arora-Sharawat and Chattopadhyay 2007; Chaudhuri et al. 2009, 2012). In addition to this, there are few negative stain EM reports that directly visualized the transition in pure and mixed spherical micelles to flexible rods on presence of salts (Imae et al. 1985; Chakraborty and Sarkar 2004). Few MD simulation studies have also been reported the transition in spherical to cylindrical SDS micelles under the effect of various salts (Sammalkorpi et al. 2009). In spite of several theoretical, computational and experimental studies on the shape transitions in self-assembled structure of amphiphiles or micelles on adding salts, there is no single report available till date that investigates how individual molecular shape of amphiphiles get affected during, or result in, this change in their macromolecular geometry. Thus here, we used three surfactants as per the classification of detergents as anionic (SDS), cationic (CTAB) and non-ionic (Tween20), to study the dynamics of individual molecular shapes of surfactants on altering the water

structure by adding three sodium salts (NaI, NaCl and Na<sub>2</sub>SO<sub>4</sub>). These electrolytes were known to affect the bulk water properties or ordering of water as NaI < NaCl < Na<sub>2</sub>SO<sub>4</sub> (Nickolov and Miller 2005; Barbosa et al. 2010). Using tryptophan octyl ester (TOE as probe)-based fluorescence assay (Arora-Sharawat and Chattopadhyay 2007), we provide evidence that shape of individual amphiphile is a key determinant which regulates the transition in the shape of micelles. Here, we propose that water structure is the most important factor compared to the ionic strength of salt and dielectric constant of solutions which alters the individual shape of amphiphiles leading to change in the shape of the micellar assembly. Based on the above, we are also able to predict the altered molecular shapes of surfactants as well as the shapes of their micellar geometry in absence and presence of these salts.

**Materials and Methods** SDS, CTAB, Tween-20, NaCl, NaI, Na<sub>2</sub>SO<sub>4</sub> and all buffers components were purchased from Sigma-Aldrich Chemicals Pvt. Ltd. TOE was purchased from Santa Cruz Biotechnology, Inc. Milli-Q water was used throughout to prepare each solution.

TOE-based fluorescence assay was performed as given in (Arora-Sharawat and Chattopadhyay 2007) and further optimized for CTAB and Tween-20. Micelles were prepared by dissolving these detergents at higher concentration than their CMC, to ensure the formation of micellar assemblies in their respective solutions. For this, we used 16 mM SDS (CMC 8 mM) (Arora-Sharawat and Chattopadhyay 2007), 8 mM CTAB (CMC 1 mM) (Neugebauer 1990) and 0.5 mM Tween-20 (CMC 0.06 mM) (Helenius et al. 1979). For REES experiments, SDS was used at 48 mM, CTAB at 8 mM and Tween-20 at 16 mM. NaCl, NaI and Na<sub>2</sub>SO<sub>4</sub> were used in 0.5 M concentration for SDS and CTAB and 1.5 M for Tween-20. The maximum molar ratio of TOE/SDS was 1:120 (mol/mol), for TOE/CTAB 1:120 (mol/mol) and for TOE/Tween-20 1:40 (mol/mol). Stock solution of TOE was dispensed to get the optimized probe to detergent ratio in each tube and dried under nitrogen purge while keeping them at ~35 °C. Then the probe was further lyophilized for ~3 h. We added desired volume of detergent and salt (to get the desired final concentration) in the dried probe and vortexed this mixture for ~3 min. The buffers used were at 5 mM; KCl-HCl (pH 1 and 2), acetate (pH 4 and 5), phosphate (pH 6), MOPS (pH 6 and 7), Tris (pH 8 and 9), CAPS (pH 10–12), Na<sub>2</sub>HPO<sub>4</sub> (pH 11.5 and 12). Blanks were prepared without TOE and controls were prepared without detergents. All samples were equilibrated at ~28 °C in the dark for 1 h. Fluorescent measurements were performed on Cary Eclipse fluorescence spectrophotometer (Agilent Technologies) using Quartz cuvette of 1 cm path length. Slit width used in all experiments was 2.5 mm except for REES and Tween-20 experiments where 5-mm-slit width was used in order to get the optimized signals. Sigmoid and Peak fittings were performed to get  $r^2 > 0.9$  using OriginPro8. Sigmoid fitting was done using dose–response function as shown below:

$$y = A1 + \frac{A2 - A1}{1 + 10^{(\log x - x)P}}$$

where  $A_2$  is  $y$  max,  $A_1$  is  $y$  min and  $P$  is hill slope.

Now, for  $[HA] \xrightleftharpoons{K} [H^+] + [A^-]$

Henderson–Hasselbalch equation:  $pH = pKa + \log\left(\frac{[A^-]}{[HA]}\right)$

Here,  $[A^-] = (y_{\max} - y)$  and  $[HA] = (y - y_{\min})$

So,

$$\begin{aligned} pH &= pKa + \log\left[\frac{(y_{\max} - y)}{(y - y_{\min})}\right] \\ \Rightarrow pH - pKa &= \log\left[\frac{(y_{\max} - y)}{(y - y_{\min})}\right] \\ \Rightarrow 10^{(pH-pKa)} &= \frac{y_{\max} - y}{y - y_{\min}} \\ \Rightarrow 10^{(pH-pKa)} &= \frac{y_{\max} - y_{\min} + y_{\min} - y}{y - y_{\min}} \\ \Rightarrow y10^{(pH-pKa)} - y_{\min}10^{(pH-pKa)} &= y_{\max} - y_{\min} + y_{\min} - y \\ \Rightarrow y10^{(pH-pKa)} + y &= y_{\min}10^{(pH-pKa)} + y_{\min} + (y_{\max} - y_{\min}) \\ \Rightarrow y[1 + 10^{(pH-pKa)}] &= y_{\min}[1 + 10^{(pH-pKa)}] + (y_{\max} - y_{\min}) \\ \Rightarrow y &= \frac{y_{\min}[1 + 10^{(pH-pKa)}] + (y_{\max} - y_{\min})}{[1 + 10^{(pH-pKa)}]} \\ \Rightarrow y &= y_{\min} + \frac{(y_{\max} - y_{\min})}{1 + 10^{(pH-pKa)}} \\ \Rightarrow y &= A_1 + \frac{(A_2 - A_1)}{1 + 10^{(pH-pka)}} \end{aligned}$$

Peak fitting was done using extreme function as given below:

$$y = y_0 + Ae^{(-e^{(-z)} - z + 1)}; \quad z = \frac{(x - xc)}{w};$$

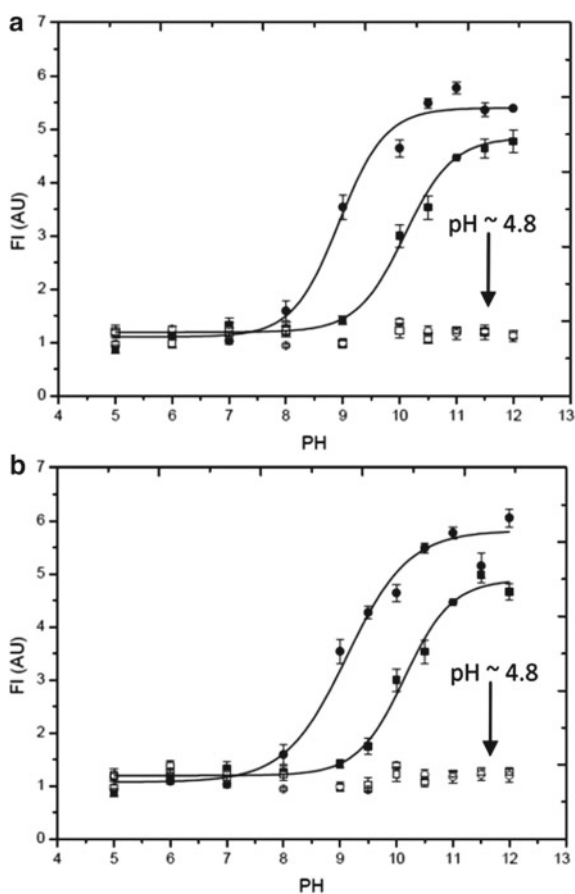
where,  $w$  is width of the curve,  $xc$  is the value at  $x$ -axis for  $y = y_{\max}$  and  $A$  is amplitude of curve.

## Results and Discussion

**TOE fluorescence as an indicator of shape transition in micelles** We first incorporated TOE in SDS micelles and monitored its fluorescence at different pH values. From the sigmoidal data obtained from pH 5 to 12 (due to the protonation and deprotonation the TOE amino group), we found pKa values of 10.1 in the absence of any salt and 8.9 (using same buffers as used in Arora-Sharawat and Chattopadhyay 2007)

and 9.1 (using different buffers at pH 6, 11.5 and 12) in presence of 0.5 M NaCl (see Fig. 8.7).

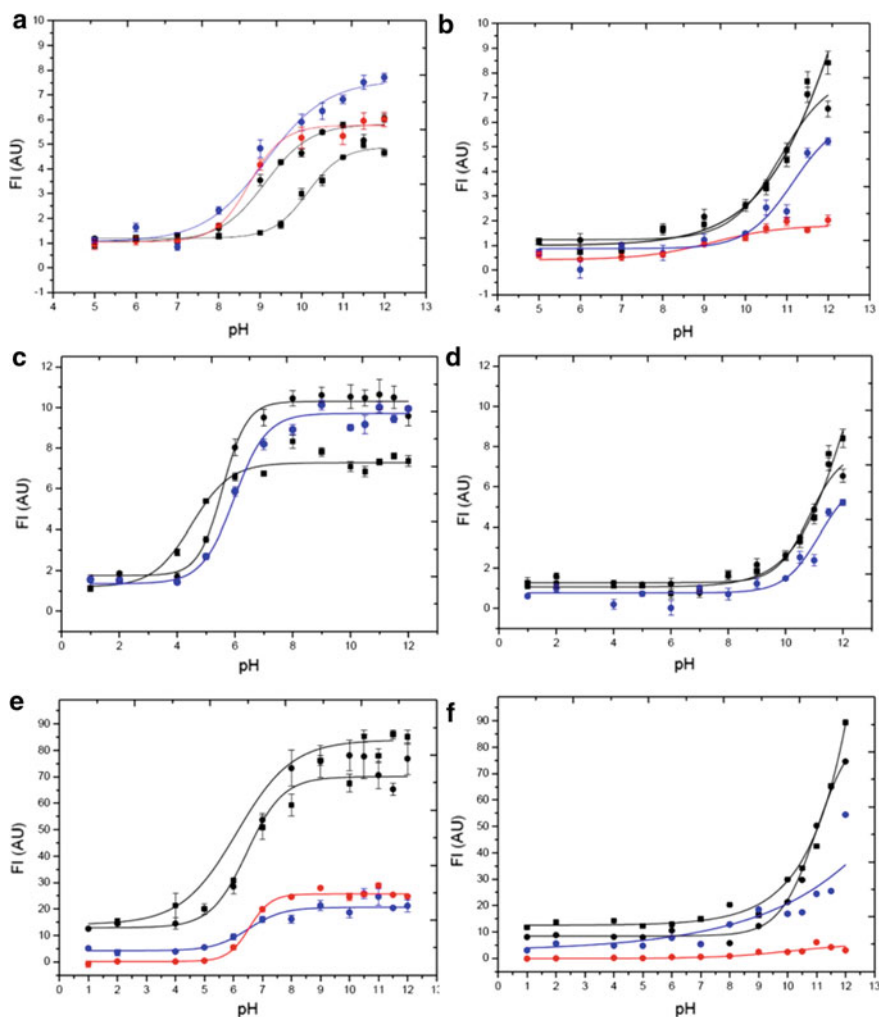
The difference of 1.2 and 1.0 unit, respectively, in pKa values was similar to the difference in pKa values reported earlier for TOE incorporated into SDS micelles under different environments. This difference is observed due to different nature of incorporation of TOE into spherical- and rod-shaped micelles—as TOE is known



**Fig. 8.7** TOE fluorescence in spherical- and rod-shaped SDS micelles. **a, b** represent mean  $\pm$  SE of fluorescence intensity of TOE in SDS micelles in the absence (filled squares) and presence (filled circles) of 0.5 M NaCl as a function of pH. In panel (**a**), buffers used in Arora-Sharawat and Chattopadhyay (2007). In panel (**b**), different buffers were used at pH 6, 11.5 and 12. The excitation wavelength was 280 nm, and emission was recorded at 340 nm. TOE/SDS ratio was 1:800 (mol/mol). The pH of all samples was lowered to  $4.8 \pm 0.1$  using 1 M acetic acid and fluorescence was shown as (empty squares) in the absence and (empty circles) presence of 0.5 M NaCl

to give environment-sensitive fluorescence signals (Arora-Sharawat and Chattopadhyay 2007). We further extended the work to test whether addition of two different salts like NaI and  $\text{Na}_2\text{SO}_4$  will also affect the pKa in similar way as NaCl. The primary objective was to explore how different liquid water structure, affected by salts/ions in the order  $\text{Na}_2\text{SO}_4 > \text{NaCl} > \text{NaI}$  or  $\text{SO}_4^{2-} > \text{Cl}^- > \text{I}^-$  (Nickolov and Miller 2005; Barbosa et al. 2010) with I-being the most disruptive (known to cause most disorderness in liquid water structure), could impact individual shape parameters of SDS (due to different hydration shell sizes around the head groups of individual molecules) thereby affecting overall geometry of the micellar self-assembly. Interestingly, we found that SDS micelles show a pKa of 9.1 with 0.5 M  $\text{Na}_2\text{SO}_4$  (similar to NaCl) and 8.7 with 0.5 M NaI (Fig. 8.8).

Therefore, in terms of micellar assembly, the order was found with SDS micelles as  $\text{NaI} > \text{NaCl} = \text{Na}_2\text{SO}_4$ . This indicates that effect of anions or salts on water structure is also playing a role in the sphere- to rod-shaped transition of SDS micelles in addition to the shielding effect of  $\text{Na}^+$  which helps in minimizing the interfacial area of SDS molecules and induces tighter and closer packing. To further understand the role of water structure on the shapes of micelles, we monitored the change in fluorescence intensity of TOE bound to CTAB micelles (CTAB is a positively charged detergent). Electron microscopy reports have suggested that CTAB aggregates to form large flexible rods of 2.2–3.0 nm radius in presence of 0.5 M NaBr (Imae et al. 1985). Thus, we treated CTAB micelles with the same concentration (0.5 M) of each salt to see their effect on its assembly process. Here, we observed a pKa of 7.3 in spherical micelles or in absence of any salt and pKa of 5.6 in 0.5 M NaCl and 6.0 in 0.5 M  $\text{Na}_2\text{SO}_4$  (see Fig. 8.8). Thus, we found the difference of 1.7 units in addition of NaCl and 1.3 units in  $\text{Na}_2\text{SO}_4$ . Interestingly, we could not take the fluorescence readings after addition of 0.5 M NaI to CTAB as this solution becomes turbid and milky white in appearance—indicating macromolecular assemblies much larger than micelles, e.g. possible long hexagonal phases or some other larger structures resulting high colloidal and turbid content. Results with CTAB indicate the order as  $\text{NaI} > \text{NaCl} > \text{Na}_2\text{SO}_4$  which is in reverse to the charge density of these anions. It is interesting to note that this order is similar (but not the same) to that obtained with SDS. Therefore, in spite of the detergents being negatively (SDS) or positively charged (CTAB), similar order of effect of anions on micellar assembly clearly suggests that shape transition in micellar assembly is majorly the consequence of perturbation in the structure of water which obviously affects the overall shape of individual surfactant molecules. A larger hydration volume of the surfactant will lead to a larger head to tail ratio, resulting in more conical individual molecular shapes which self-assembled into spherical micelles. Subsequently with decreasing head size, individual molecular shapes start becoming cylindrical leading to cylindrical micellar assemblies. Having obtained interesting insights by using negatively and positively charged detergents, we further investigated the (possible) shape transitions in non-ionic detergent, Tween-20. It is known that only high concentrations of salts may affect the CMC and/or the assembly of non-ionic detergents (Ray and Nemethy 1971). Thus, we used 1.5 M concentration for each salt as maximum solubility of  $\text{Na}_2\text{SO}_4$  is 1.5 M at 28 °C. Here, we observed pKa of 7.1 without any salt (spherical micelles) which decreases



**Fig. 8.8** TOE fluorescence in spherical- and rod-shaped micellar assemblies. Panels **a, c, e** represent mean  $\pm$  SE of fluorescence intensity of TOE in micelles in the absence (squares) and presence (circles) of salts as a function of pH. Panels **b, d, f** represent mean  $\pm$  SE of fluorescence intensity of TOE in controls (without micelles) in the absence (squares) and presence (circles) of salts as a function of pH. Panels **a, b** represent SDS micelles, **c, d** used for CTAB and **e, f** for Tween-20 micelles. Black circles were used for NaCl, blue circles for  $\text{Na}_2\text{SO}_4$  and red circles were used for NaI. The excitation wavelength was 280 nm, and emission was recorded at 340 nm. TOE/SDS ratio was 1:800, TOE/CTAB ratio was 1:400 and TOE/Tween-20 was 1:80 (mol/mol)



to 6.5 in case of each of the three salts (Fig. 8.8). Remarkably, we observed similar pKa and a difference of 0.6 units in case of all three salts. So, the order found here is  $\text{NaI} = \text{NaCl} = \text{Na}_2\text{SO}_4$ . Here, it is important to note that as compared to the SDS and CTAB, Tween-20 contains a large bulky head group to be with. Clearly, this large head group seems to limit the effect of salts since the changes in hydration volume of the head group (due to changed water structure) are negligible compared to original head size. Therefore, the individual molecular shape of Tween-20 remains same in all three salts as conical resulting in only spherical micelles. Therefore, our results clearly show that transition in the molecular shapes of amphiphiles is the root cause of the shape transitions in micellar assembly. However, this change in shape parameter is allowed within the constraints of its chemical structure and in turn facilitated by the variation in water structure in aqueous medium.

**Red edge excitation shift of TOE in spherical and cylindrical micelles** Red edge excitation shift (REES) is defined as the shift of emission spectra towards longer wavelength on increasing the excitation wavelength. This effect would be seen if fluorophores occupy motionally confined regions such as interfacial regions of micelles as occupied by TOE (Arora-Sharawat and Chattopadhyay 2007). Thus, REES measurements on TOE would also provide an assay to monitor the transition of micellar geometries. Therefore, in order to confirm our above findings based on pKa values, we performed REES experiments on SDS, CTAB and Tween20 micelles. In case of SDS, REES experiments were conducted at pH 5 and 11 only as TOE predominately presents in protonated form at pH 5 and deprotonated forms exist at pH 11. Similarly, for CTAB, REES experiments were performed at pH 2 and pH 11. In case of Tween-20, pH 4 and pH 11 were used for REES experiments (Fig. 8.8). We observed a red shift of 12 nm ( $336 \pm 0.03$  nm to  $348 \pm 0.09$  nm) in the maximum fluorescence emission of TOE in SDS micelles with increasing pH from 5 to 11 due to the deprotonation of TOE. As expected, we also observed a similar blue shift of ~2 nm (336 nm to 334) in fluorescence emission maxima at pH 5 and from 348 nm to 346 nm at pH 11 upon addition of 0.5 M NaCl (see comparative Table 8.1).

These results indicate that the shape transition of the micelles was accompanied by the decrease in polarity around TOE (Arora-Sharawat and Chattopadhyay 2007). Interestingly, we observed similar blue shift with all three salts. Figure 8.9 shows fluorescence emission spectra of TOE in SDS micelles at pH 5 and pH 11 in absence and presence of 0.5 M NaCl,  $\text{Na}_2\text{SO}_4$  and NaI.

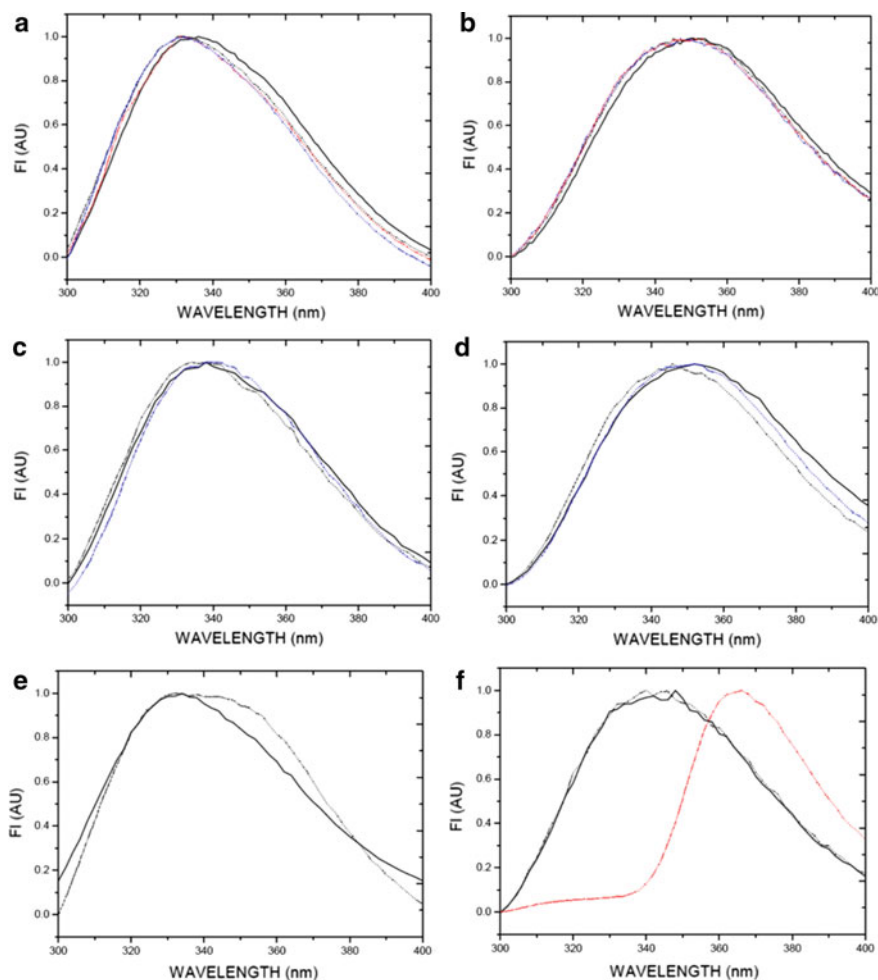
We observed a shift in emission maximum from 336 to 341 nm and REES of 5 nm at pH 5 in spherical micelles (Fig. 8.10).

We also observed 348–351 nm shift and REES of 3 nm at pH 11 in spherical micelles. After adding 0.5 M NaCl, we found 334–340 nm shift and REES of 6 nm at pH 5 and 346–350 nm shift with REES of 4 nm at pH 11 in rod-shaped SDS micelles. A comparative Table 8.2 was inserted to compare our observation with the previously reported work. Again with 0.5 M  $\text{Na}_2\text{SO}_4$ , emission shift was found from 334 to 340 nm, i.e. REES of 6 nm at pH 5 and 346–350 nm with REES of 4 nm at pH 11. We observed highest shift from 334 to 342 nm, i.e. REES of 8 nm

**Table 8.1** Fluorescence emission maximum of TOE in SDS, CTAB and Tween-20 micelles at 280 nm

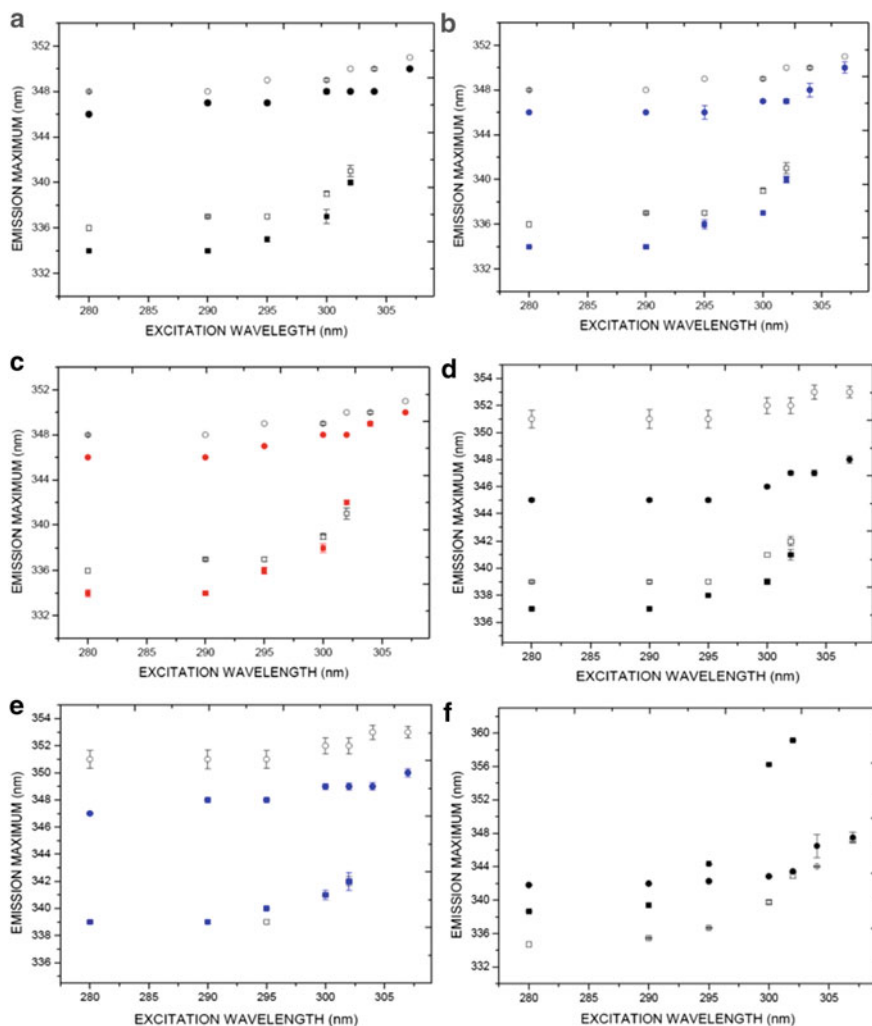
Host	Fluorescence emission maximum from Arora-Sharawat and Chattopadhyay (2007) (nm)	Fluorescence emission maximum from this work (nm) [mean $\pm$ SE]
<i>(A) pH 5</i>		
SDS micelles	335	336 $\pm$ 0.03
SDS micelles (in presence of 0.5 M NaCl)	332	334 $\pm$ 0.1
SDS micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	334 $\pm$ 0.07
SDS micelles (in presence of 0.5 M NaI)	NA	334 $\pm$ 0.3
<i>(B) pH 11</i>		
SDS micelles	345	348 $\pm$ 0.09
SDS micelles (in presence of 0.5 M NaCl)	343	346 $\pm$ 0.09
SDS micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	346
SDS micelles (in presence of 0.5 M NaI)	NA	346 $\pm$ 0.1
<i>(A) pH 2</i>		
CTAB micelles	NA	339 $\pm$ 0.06
CTAB micelles (in presence of 0.5 M NaCl)	NA	337 $\pm$ 0.08
CTAB micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	339 $\pm$ 0.1
<i>(B) pH 11</i>		
CTAB micelles	NA	351 $\pm$ 0.66
CTAB micelles (in presence of 0.5 M NaCl)	NA	345 $\pm$ 0.13
CTAB micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	347 $\pm$ 0.04
<i>(A) pH 4</i>		
Tween-20 micelles	NA	335
Tween-20 micelles (in presence of 1.5 M NaCl)	NA	339
<i>(A) pH 11</i>		
Tween-20 micelles	NA	342 $\pm$ 0.03
Tween-20 micelles (in presence of 1.5 M NaCl)	NA	342 $\pm$ 0.03

NA stands for not available



**Fig. 8.9** Fluorescence emission spectra of TOE in micelles. Panels **a–f** represent emission spectra of TOE bound to micelles in the absence (solid line) and presence (dotted line) of salts. Panels **a, b** were used for SDS micelles at pH 5 and pH 11, respectively. Panels **c, d** represent emission spectra of TOE bound to CTAB micelles at pH 2 and 11, respectively. Panels **e, f** represent Tween-20 micelles at pH 4 and 11, respectively. Black colour was used for NaCl, blue for Na<sub>2</sub>SO<sub>4</sub> and red was used for NaI. The excitation wavelength was 280 nm. TOE/SDS ratio was 1:120 at pH 5 and 1:160 at pH 11. TOE/CTAB ratio was 1:120 at pH 5 and 1:160 at pH 11. TOE/Tween-20 ratio was 1:40 at pH 4 and 1:60 (mol/mol) at pH 11

at pH 5 and a shift of 346–350 nm with REES 4 nm at pH 11 in 0.5 M NaI or in rod-shaped micelles. The presence of higher REES in case of NaI at pH 5 indicates the formation of most tightly packed rods in present of NaI. Thus, REES experiments also suggest that the effect of ordering on shape transition of SDS micelles, i.e. NaI >



**Fig. 8.10** Change in emission maximum on increasing excitation wavelength of TOE in micelles. **a–c** represent emission maximum in SDS micelles at pH 5 in the absence (empty squares) and presence (filled squares) of salt, and pH 11 in the absence or spherical micelle (empty circles), the presence of salt or rod-shaped (filled circles) micelles. Panel **d**, **e** represent emission maximum in CTAB micelles at pH 2 and pH 11. Panel **f** represents emission maximum in Tween-20 micelles at pH 4 and pH 11. Black coloured filled symbols represent NaCl, blue colour was used for Na<sub>2</sub>SO<sub>4</sub> and red represents Na. TOE/SDS ratio was 1:120 at pH 5 and 1:160 at pH 11. TOE/CTAB ratio was 1:120 at pH 5 and 1:160 at pH 11. TOE/Tween-20 ratio was 1:40 at pH 4 and 1:60 (mol/mol) at pH 11

**Table 8.2** Red edge excitation shift in TOE bound to SDS, CTAB and Tween-20 micelles

Host	REES reported in Arora-Sharawat and Chattopadhyay (2007) (nm)	REES observed from our work (nm)
<i>(A) pH 5</i>		
SDS micelles	4	5
SDS micelles (in presence of 0.5 M NaCl)	6	6
SDS micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	6
SDS micelles (in presence of 0.5 M NaI)	NA	8
<i>(B) pH 11</i>		
SDS micelles	4	3
SDS micelles (in presence of 0.5 M NaCl)	5	4
SDS micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	4
SDS micelles (in presence of 0.5 M NaI)	NA	4
<i>(A) pH 2</i>		
CTAB micelles	NA	3
CTAB micelles (in presence of 0.5 M NaCl)	NA	4
CTAB micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	3
<i>(B) pH 11</i>		
CTAB micelles	NA	2
CTAB micelles (in presence of 0.5 M NaCl)	NA	3
CTAB micelles (in presence of 0.5 M Na <sub>2</sub> SO <sub>4</sub> )	NA	3
<i>(A) pH 4</i>		
Tween-20 micelles	NA	8
Tween-20 micelles (in presence of 0.5 M NaCl)	NA	20
<i>(B) pH 11</i>		
Tween-20 micelles	NA	5
Tween-20 micelles (in presence of 0.5 M NaCl)	NA	5

NA stands for not available

$\text{NaCl} = \text{Na}_2\text{SO}_4$ , as suggested by the difference in pKa values of TOE fluorescence in micellar assemblies.

Figure 8.9 depicts fluorescence spectra of CTAB bound TOE at pH 2 and 11 in absence and presence of 0.5 M NaCl, and  $\text{Na}_2\text{SO}_4$ . We observed a red shift of 12 nm in the maximum fluorescence emission of TOE in CTAB micelles (from 339 to 351 nm) upon increasing the pH from 2 to 11 due to the deprotonation of TOE (see Table 8.1). We observed a blue shift of 2–6 nm in the maximum of fluorescence emission (from 339 to 337 nm, and 351 to 345 nm) upon addition of 0.5 M NaCl at both pH which indicates decrease in polarity around micelle bound TOE due to the sphere- to rod-shaped transition of the CTAB micelles (see Table 8.1). We observed a blue shift of 4 nm in the maximum of fluorescence emission (from 351 to 347 nm) upon addition of 0.5 M  $\text{Na}_2\text{SO}_4$  at pH 11. However at pH 2, no blue shift was observed in case of  $\text{Na}_2\text{SO}_4$ . We observed 339 to 342 nm shift in emission maximum and REES of 3 nm at pH 2 in spherical micelles (see Fig. 8.10 and Table 8.2). At pH 11, a shift from 351 to 353 nm was found with REES of 2 nm in spherical micelles. We found 337 to 341 nm shift and REES of 4 nm at pH 2 in 0.5 M NaCl or in rod-shaped micelles. We observed 345–348 nm shift and REES of 3 nm at pH 11 in 0.5 M NaCl or in rod-shaped micelles. In case of 0.5 M  $\text{Na}_2\text{SO}_4$ , we observed 339–342 nm shift, i.e. REES of 3 nm at pH 2, and 347–350 nm shift with REES of 3 nm at pH 11. Thus, we found increased REES in case of NaCl as compared to  $\text{Na}_2\text{SO}_4$  which indicates tighter packing of surfactants in presence of NaCl than  $\text{Na}_2\text{SO}_4$ . We could not perform REES experiments on NaI because of the formation of insoluble particles in milky white solution. Thus, REES values also suggest that the effect of ordering on shape transition of CTAB micelles is  $\text{NaI} > \text{NaCl} > \text{Na}_2\text{SO}_4$ .

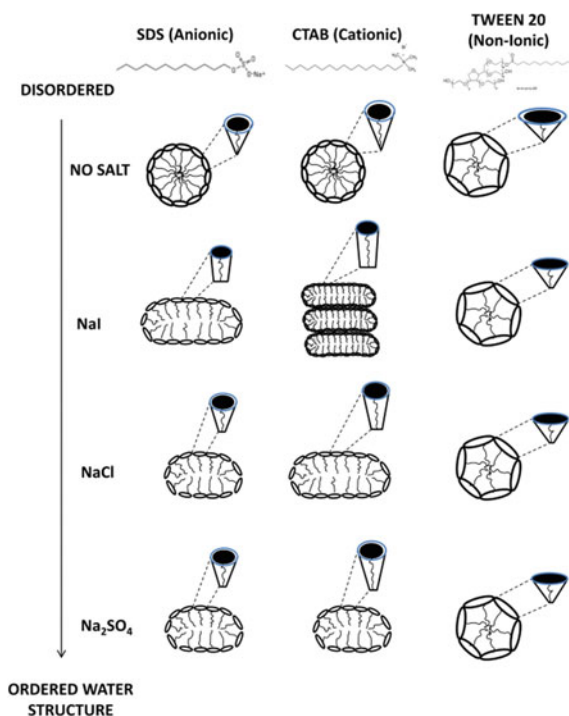
Figure 8.8 represents fluorescence emission spectra of TOE in Tween-20 micelles at pH 4 and pH 11 in absence and presence of 1.5 M NaCl and NaI. We observed a red shift of 7 nm in the maximum fluorescence emission of TOE in Tween-20 micelles (from 335 to 342 nm) upon increasing the pH from 4 to 11, indicative of deprotonation of TOE (Table 8.1). There was no blue shift observed in addition of 1.5 M NaCl, and we could not capture these signals in case of 1.5 M NaI and  $\text{Na}_2\text{SO}_4$  also. Thus, REES experiments did not provide any insights into Tween-20 micellar assembly in different salts. That said, it can also be interpreted that Tween-20 micellar assembly is not, or equally affected, by different salts—a result that is in agreement with our earlier findings using pKa values.

### **Predicted possible molecular and assembled micellar shapes of surfactants**

Above observations clearly support that the shape transition from spherical to cylindrical micelles is largely govern by the molecular shape of surfactants under the effect of altered water structures. Extent to which these molecular shapes can change seems to be dependent upon the degree of perturbation in the structure of liquid water. More compact rods were observed in case of salt causing less ordering in water structure. However, chemical structure of each amphiphiles itself creates some constraints up to which maximum transition in these molecular shapes are permissible. Here based

on these inferences, we predict the possible change in molecular shapes and assembled micellar structure in the absence and presence of three different electrolytes (Fig. 8.11).

Israelachvili et al. (1976) gave a concept of shape parameter which defines the individual shapes of the amphipathic lipid molecules as  $S = v/al$ , where “ $v$ ” represents hydrophobic tail volume of the lipid, “ $a$ ” is the optimum interfacial head area and  $l$  is the maximum tail length of the amphiphiles. It was also proposed that this packing parameter or shape parameter can determine the phase of different lipids such as for  $S < 1/3$  will form spherical micelles,  $1/3 < S < 1/2$  will form non-spherical micelles,  $1/2 < S < 1$  form bilayers and  $S > 1$  will form inverted structures. Tanford (1973) and (1978) also proposed geometrical considerations and theoretical equations to compute for the maximum extended length of the lipid and core volume of spherical micelles or hydrophobic volume of individual lipid from their chemical structures as:



**Fig. 8.11** Predicted possible shapes of individual surfactants and micellar assemblies. Figure illustrates the possible coupling between the effect of electrolytes on water structures and shape transition in micellar assembly. NaI, which tends to induce more disordering as compared to NaCl and Na<sub>2</sub>SO<sub>4</sub>, causes maximum reduction in the hydrodynamic radius around the interfacial head region of SDS and CTAB. This further induces more cylindrical geometry and thus leads to the formation of longer rods in presence of NaI. However, this effect was completely absent in Tween-20 due to the steric hindrance produced by its bulky head group

$l = 1.5 + 1.265 n'c$ , where  $l$  is the extended length of the hydrophobic tail in Å and  $n'c$  is the number of embedded carbons in the core of micelle.

$v = (27.4 + 26.9 n'c)m'$ , where  $v$  is the core volume of micelle in (Å)<sup>3</sup>,  $n'c$  is the number of embedded carbons in the core of micelle and  $m'$  is the number of hydrocarbon chain per micelle (or equal to aggregation number in single-tailed surfactants). However, “ $a$ ” is interfacial area between head and tail region and depends on the physico-chemical conditions. We used these equations and computed the volume and length for spherical micelles of SDS, CTAB and Tween-20 on the basis of their chemical structures and other available information. SDS and CTAB have minor differences in their extended lengths (16.7 Å and 21.7 Å, respectively) and hydrophobic tail volumes (350.2 Å<sup>3</sup> and 457.8 Å<sup>3</sup>, respectively) and thus should have similar  $S$ -values which are also indicated from their similar aggregation numbers ~62 (Arora-Sharawat and Chattopadhyay 2007) and ~60 (Neugebauer 1990), respectively. On the other hand, Tween-20 consists of a large bulky head group and tail length of 12 carbons (tail length is similar to SDS i.e. 16.7 Å). Therefore, Tween-20 has high interfacial surface area which results into a wide-shaped cone geometry that cannot be accommodated in large numbers to form spherical micelles with aggregation number ~22 (Acharya et al. 1997). Addition of electrolytes in all three detergents decrease the interfacial surface area of each molecule which facilitate the change in individual molecular shape from conical to upright truncated cone and thus, further results cylindrical micelles. However, permissible range up to which change in individual molecular shape is possible, actually conferred by its own chemical structure or geometrical constraints under the influence of hydrophobic effect. It is known that these salts affect the structure of water as Na<sub>2</sub>SO<sub>4</sub> > NaCl > NaI (Nickolov and Miller 2005; Barbosa et al. 2010). In case of charged micelles (SDS and CTAB), NaI shows its maximum effect in allowing the formation of long compact cylindrical rods and even form much bigger particles showing hexagonal phase formation in case of CTAB. This indicates that NaI induces maximum reduction in the interfacial surface area of individual surfactant and thus aids maximum transition in their molecular shapes to form almost cylindrical-shaped molecules, which in turn, favour the formation of long compact rods. However, this effect is absent in case of non-ionic detergent (Tween-20) because of its substantially bulkier head group which limits the decrease in its effective surface area and thus, predominantly conical individual molecules result in formation of spherical micelles in presence of any salt. Additionally, shape transition in micelles known to be closely coupled with the drastic decrease in CMC of micelles on addition of electrolytes (Ray and Nemethy 1971; Imae et al. 1985; Rawat and Chattopadhyay 1999; Chakraborty and Sarkar 2004; Ericsson et al. 2004; Arora-Sharawat and Chattopadhyay 2007; Chaudhuri et al. 2009, 2012). For example; CMC of SDS (anionic) is 8.0 mM which drops to 0.5 mM in presence of 0.5 M NaCl (Tanford 1973). However, non-ionic surfactants (usually have bulky head group which provide lot of steric hindrance) require high concentration of salts as compared to charged surfactants to cause significant changes in their CMC and/or formation of cylindrical micelles (Ray and Nemethy 1971). Conceptually, shape parameter is also defined as a post-assembly parameter; i.e. shape parameter can be computed only after the formation of an assembled geometry. Our study reports



an increase in the  $S$  value of surfactants (due to decrease in the interfacial area of surfactants in the presence of salts) accompanied their transition from globular to rod-shaped micelles. Thus, CMC and shape parameter ( $S$ ) seem to share an inverse relationship and act as two indicators to monitor the change in micellar geometry. Our findings clearly provide evidence that individual molecular shape of amphiphiles that are primarily responsible for different geometries of hydrophobic effect-driven self assemblies, can be altered by changing the structure of water. Interestingly, recent studies reported that the restricted non-random dipolar arrangement of water molecules in cylindrical micelles led to substantial changes in the dipole potential of micelles. Similar to our results, it has been reported that this change does not depend on the charge of surfactants and was absent in non-ionic detergents (Sarkar and Chattopadhyay 2015, 2016). This clearly supports that structure and orientation of water molecules around the amphiphiles play a crucial role in facilitating the transition in their macromolecular assembly. Thus, this work provides a strong direction towards further studies for creating/controlling geometrical aspects of hydrophobic effect-driven macromolecular self-assembled structures by varying water structure.

**Conclusions** Membrane structures undergo continuous rearrangements or remodelling during various cellular processes. Enormous diversity and distribution of these membrane lipids maintain the curvature and functionalities in various membranous organelles. It has been shown that the asymmetric distribution of individual molecular shapes of lipids could induce curvature formation in membranes. Preference of individual molecular shapes to form a particular phase in aqueous solutions has been investigated using computational, theoretical and experimental approaches. Micelles have served as a widely studied experimental system to understand the self-assembly of amphiphiles driven by hydrophobic effect. Structure of water, that provides the aqueous medium for micellar assemblies, has been known to be altered by addition of electrolytes. While addition of these salts had been reported to induce sphere to cylindrical micelle formation in various surfactants, its impact on individual molecular shapes of the self-assembling surfactants that leads to formation of the final micellar geometry was unclear prior to this work. Thus, in order to study the dynamics of molecular shapes of surfactants accompanying the sphere- to rod-shaped transition in micelles, we used two charged: SDS (anionic), CTAB (cationic) and one non-ionic (Tween-20) detergent. Three electrolytes were used to disturb the water structure by increasing the orderliness in liquid water. In spite of common  $\text{Na}^+$  ions in all three salts, we observed difference in the  $\text{pK}_a$  values of TOE bound to SDS micelles in case of each salt. This indicates that the hydrophobic effect plays a major role as compared to the shielding effect of  $\text{Na}^+$  ions on negative charge of SDS. In CTAB bound TOE, we found maximum difference in  $\text{pK}_a$  in NaI (similar to SDS) than NaCl and  $\text{Na}_2\text{SO}_4$ . This suggests that the charge on the micellar assembly is not so critical. However, this is structure of water which drives the modification in the individual molecular shapes that further, act as the key determinant in controlling the self-assembly of micellar structures. This inference was also supported by Tween-20 where all three salts showed equal  $\text{pK}_a$  values because of the large bulky head group of Tween-20. Thus, these findings provide evidence that the molecular shapes provide the first

constraint and act as a primary factor in controlling the self-assembly of amphiphiles. These molecular shapes can be altered by perturbing the water structure; however, more disordering or lesser disturbance in bulk water structure yields more compact long rods (as in case of NaI) as compared to the water with more ordered structures (with NaCl and Na<sub>2</sub>SO<sub>4</sub>). Based on the above findings, we can now also predict the possible changes in individual molecular shapes of surfactants and thus their impact on the shape of overall assembled micelles in different aqueous conditions.

## References

- Acharya KR, Bhattacharya SC, Moulik SP (1997) The surfactant concentration-dependent behaviour of safranin T in Tween (20, 40, 60, 80) and Triton X-100 micellar media. *J Photochem Photobiol A Chem* 109:29–34
- Agozzino L, Dill KA (2018) Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc Natl Acad Sci USA* 115:9092–9097
- Andreyev AY, Fahy E, Guan Z, Kelly S, Li X, McDonald JG, Milne S, Myers D, Park H, Ryan A, Thompson BM, Wang E, Zhao Y, Alex Brown H, Merrill AH, Raetz CRH, Russell DW, Subramaniam S, Dennis EA (2010) Subcellular organelle lipidomics in TLR-4-activated macrophages. *J Lipid Res* 51:2785–2797
- Arora-Sharawat A, Chattopadhyay A (2007) Effect of structural transition of the host assembly on dynamics of a membrane-bound tryptophan analogue. *Biophys Chem* 129:172–180
- Baldwin RL (1996) How Hofmeister ion interactions affect protein stability. *Biophys J* 71:2056–2063
- Bangham AD (1972) Model membranes. *Chem Phys Lipids* 8:386–392
- Bansal S, Mittal A (2013) Extracting curvature preferences of lipids assembled in flat bilayers shows possible kinetic windows for genesis of bilayer asymmetry and domain formation in biological membranes. *J Membrane Biol* 246:557–570
- Bansal S, Mittal A (2015) A statistical anomaly indicates symbiotic origins of eukaryotic membranes. *Mol Biol Cell* 26:1238–1248
- Barbosa AM, Santos IJB, Ferreira GMD, da Silva MDH, Teixeira A, da Silva LHM (2010) Microcalorimetric and SAXS determination of PEO-SDS interactions: the effect of cosolutes formed by ions. *J Phys Chem B* 114:11967–11974
- Brini E, Fennell CJ, Fernandez-Serra M, Hribar-Lee B, Lukšić M, Dill KA (2017) How Water's properties are encoded in its molecular structure and energies. *Chem Rev* 117:12385–12414
- Brown DA, London E (1998) Functions of lipid rafts in biological membranes. *Annu Rev Cell Dev Biol* 14:111–136
- Chakraborty H, Sarkar M (2004) Optical spectroscopic and TEM studies of cationic micelles of CTAB/SDS and their interaction with a NSAID. *Langmuir* 20:3551–3558
- Chaudhuri A, Haldar S, Chattopadhyay A (2009) Organization and dynamics in micellar structural transition monitored by pyrene fluorescence. *Biochem Biophys Res Commun* 390:728–732
- Chaudhuri A, Haldar S, Chattopadhyay A (2012) Structural transition in micelles: novel insight into microenvironmental changes in polarity and dynamics. *Chem Phys Lipids* 165:497–504
- Chernomordik LV, Kozlov MM (2003) Protein-lipid interplay in fusion and fission of biological membranes. *Annu Rev Biochem* 72:175–207
- Chernomordik LV, Kozlov MM, Melikyan GB, Abidor IG, Markin VS, Chizmadzhev YA (1985) The shape of lipid molecules and monolayer membrane fusion. *Biochim Biophys Acta* 812:643–655
- Chernomordik L, Leikina E, Cho M-S, Zimmerberg J (1995) Control of baculovirus gp64-induced syncytium formation by membrane lipid composition. *J Virol* 69:3049–3058

- Christian DA, Tian A, Ellenbroek WG, Levental I, Rajagopal K, Janmey PA, Liu AJ, Baumgart T, Discher DE (2009) Spotted vesicles, striped micelles and Janus assemblies induced by ligand binding. *Nat Mater* 8:843–849
- Christiansson A, Kuypers FA, Roelofsen B, Op Den Kamp JAF, Van Deenen LLM (1985) Lipid molecular shape affects erythrocyte morphology: a study involving replacement of native phosphatidylcholine with different species followed by treatment of cells with sphingomyelinase C or phospholipase A2. *J Cell Biol* 101:1455–1462
- Collins KD (1997) Charge density-dependent strength of hydration and biological structure. *Biophys J* 72:65–76
- Collins KD (2004) Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process. *Methods* 34:300–311
- Cooke IR, Deserno M (2006) Coupling between lipid shape and membrane curvature. *Biophys J* 91:487–495
- Deamer DW (2019) *Assembling life: how can life begin on Earth and other habitable planets?* Oxford University Press
- Dill KA, Bromberg S (2011) *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*, 2nd edn. Garland Science, New York
- Dill KA, Truskett T, Vlachy V, Hribar-Lee B (2005) Modeling water, the hydrophobic effect, and ion solvation. *Ann Rev Biophys Biomol Struct* 34:173–199
- Dill KA, Ghosh K, Schmit JD (2011) Physical limits of cells and proteomes. *Proc Natl Acad Sci USA* 108:17876–17882
- Eisenberg D, Kauzmann W (1969) *The structure and properties of water*. Academic Press, Oxford University Press, Oxford
- Ericsson CA, Soderman O, Garamus VM, Bergstrom M, Ulvenlund S (2004) Effects of temperature, salt, and deuterium oxide on the self-aggregation of alkylglycosides in dilute solution. 1. n-nonyl- $\beta$ -D-glucoside. *Langmuir* 20:1401–1408
- Fennell CJ, Kehoe C, Dill KA (2010) Oil/water transfer is partly driven by molecular shape, not just size. *J Am Chem Soc* 132:234–240
- Ghosh K, Dill KA (2010) Cellular proteomes have broad distributions of protein stability. *Biophys J* 99:3996–4002
- Ghosh K, de Graff AMR, Sawle L, Dill KA (2016) Role of proteome physical chemistry in cell behavior. *J Phys Chem B* 120:9549–9563
- Heinonem JK (2001) *Biological role of inorganic phosphate*. Springer Science, Business Media, LLC, Berlin
- Helenius A, Mccaslin DR, Fries E, Tanford C (1979) Properties of detergents. *Meth Enzymol* 56:734–749
- Imae T, Kamiya R, Ikeda S (1985) Formation of spherical and rod-like micelles of cetyltrimethylammonium bromide in aqueous NaBr solutions. *J Colloid Interface Sci* 108:215–225
- Israelachvili JN, Mitchell DJ, Ninham BWJ (1976) Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers. *Chem Soc Faraday Trans* 72(2):1525–1568
- Lahtee PJ, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, Nielsen J (2017) Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst* 4:495–504
- Mahler J, Ingmar P (2012) A study of the hydration of the alkali metal ions in aqueous solution. *Inorg Chem* 51:425–438
- Maitra A, Dill KA (2015) Bacterial growth laws reflect the evolutionary importance of energy efficiency. *Proc Natl Acad Sci USA* 112:406–411
- Marcus Y (2009) Effect of ions on the structure of water: structure making and breaking. *Chem Rev* 109:1346–1370
- McMahon HT, Gallop JL (2005) Membrane curvature and mechanisms of dynamic cell membrane remodelling. *Nature* 438:590–596
- Mittal A, Grover RJ (2010) Self-assembly of biological membranes into 200–400 nm aqueous compartments. *Nanosci Nanotechnol* 10:3085–3090

- Mittal A, Jayaram B (2011) The newest view on protein folding: stoichiometric and spatial unity in structural and functional diversity. *J Biomol Struct Dyn* 28:669–674
- Mittal A, Singh S (2018) Insights into eukaryotic evolution from transmembrane domain lengths. *J Biomol Struct Dyn* 36:2194–2200
- Mittal A, Leikina E, Bentz J, Chernomordik LV (2002) Kinetics of influenza hemagglutinin-mediated membrane fusion as a function of technique. *Anal Biochem* 30:145–152
- Mukherjee S, Soe TT, Maxfield FR (1999) Endocytic sorting of lipid analogues differing solely in the chemistry of their hydrophobic tails. *J Cell Biol* 144:1271–1284
- Naresh M, Hasija V, Sharma M, Mittal A (2010) Synthesis of cellular organelles containing nanomagnets stunts growth of magnetotactic bacteria. *J Nanosci Nanotechnol* 10:4135–4144
- Naresh M, Das S, Mishra P, Mittal A (2012) The chemical formula of a magnetotactic bacterium. *Biotechnol Bioeng* 109:1205–1216
- Nelson DL, Cox MM (2008) *Lehninger principles of biochemistry*, 5th edn. W.H. Freeman & Co Ltd.
- Neugebauer JM (1990) Detergents: an overview. *Meth Enzymol* 182:239–253
- Nicolov ZS, Miller JDJ (2005) Water structure in aqueous solutions of alkali halide salts: FTIR spectroscopy of the OD stretching band. *Colloid Interface Sci* 287:572–580
- Poos MI, Costello R, Carlson-Newberry SJ (1999) Institute of Medicine (US) committee on military nutrition research, the role of protein and amino acids in sustaining and enhancing performance. National Academies Press, Washington, DC. <https://doi.org/10.17226/9620>
- Rawat SS, Chattopadhyay A (1999) Structural transition in the micellar assembly: a fluorescence study. *J Fluoresc* 9:233–244
- Ray A, Nemethy G (1971) Effects of ionic protein denaturants on micelle formation by nonionic detergents. *J Am Chem Soc* 93:6787–6793
- Roux A, Cuvelier D, Nassoy P, Prost J, Bassereau P, Goud B (2005) Role of curvature and phase transition in lipid sorting and fission of membrane tubules. *EMBO J* 24:1537–1545
- Sammalkorpi M, Karttunen M, Haataja M (2009) Ionic surfactant aggregates in saline solutions: sodium dodecyl sulfate (SDS) in the presence of excess sodium chloride (NaCl) or calcium chloride (CaCl<sub>2</sub>). *J Phys Chem B* 113:5863–5870
- Santra M, Farrell DW, Dill KA (2017) Bacterial proteostasis balances energy and chaperone utilization efficiently. *Proc Natl Acad Sci USA* 114:E2654–E2661
- Sarkar P, Chattopadhyay A (2015) Dipolar rearrangement during micellization explored using a potential-sensitive fluorescent probe. *Chem Phys Lipids* 191:91–95
- Sarkar P, Chattopadhyay A (2016) Micellar dipole potential is sensitive to sphere-to-rod transition. *Chem Phys Lipids* 195:34–38
- Schafer LV, Marrink SJ (2010) Partitioning of lipids at domain boundaries in model membranes. *Biophys J* 99:L91–L93
- Sezonov G, Joseleau-Petit D, D'Ari R (2007) *Escherichia coli* physiology in Luria-Bertani broth. *J Bacteriol* 23:8746–8749
- Silverstein KAT, Haymet ADJ, Dill KA (1998) A simple model of water and the hydrophobic effect. *J Am Chem Soc* 120:3166–3175
- Singer SJ, Nicolson GL (1972) The fluid mosaic model of the structure of cell membranes. *Science* 175:720–731
- Singh S, Mittal A (2016) Transmembrane domain lengths serve as signatures of organismal complexity and viral transport mechanisms. *Sci Rep* 6:22352. <https://doi.org/10.1038/srep22352>
- Singh S, Ponnappan N, Verma A, Mittal A (2019) Osmotic tolerance of avian erythrocytes to complete hemolysis in solute free water. *Sci Rep* 9:7976. <https://doi.org/10.1038/s41598-019-44487-7>
- Southall NT, Dill KA, Haymet ADJ (2002) A view of the hydrophobic effect. *J Phys Chem B* 106:521–533

- St Vincent MR, Colpitts CC, Ustinov AV, Muqadas M, Joyceb MA, Barsby NL, Epanand RF, Epanand RM, Khramyshev SA, Valueva OA, Korshun VA, Tyrrell DLJ, Schang LM (2010) Rigid amphipathic fusion inhibitors, small molecule antiviral compounds against enveloped viruses. *PNAS* 107:17339–17344
- Szenk M, Dill KA, de Graff AMR (2017) Why do fast-growing bacteria enter overflow metabolism? Testing the membrane real estate hypothesis. *Cell Syst* 5:95–104
- Tanford C (1973) *The hydrophobic effect: formation of micelles and biological membranes*, 1st edn. Academic Press, Wiley-Interscience, New York
- Tanford C (1978) The hydrophobic effect and the organization of living matter. *Science* 200:1012–1018
- Truskett TM, Dill KA (2003) A simple analytical model of water. *Biophys Chem* 105:449–459
- Urbic T, Dill KA (2017) Analytical theory of the hydrophobic effect of solutes in water. *Phys Rev E* 96:32101
- Urbic T, Dill KA (2018) Water Is a Cagey Liquid. *J Am Chem Soc* 140:17106–17113
- Urbic T, Vlachy V, Kalyuzhnyi YV, Dill KA (2007) Theory for the solvation of nonpolar solutes in water. *J Chem Phys* 127:174505
- Van Meer G, Voelker DR, Feigenson GW (2008) Membrane lipids: where they are and how they behave. *Nat Rev Mol Cell Biol* 9:112–124
- Voet D, Voet JG, Pratt CW (2013) *Principles of biochemistry*. Wiley, New York
- Wagoner J, Dill KA (2019) Mechanisms for achieving high speed and efficiency in biomolecular machines. *Proc Natl Acad Sci USA* 116:5902–5907
- Wetlaufer DB, Malik SK, Stoller L, Coffin RL (1964) Nonpolar group participation in the denaturation of proteins by urea and guanidinium salts. Model compound studies. *J Am Chem Soc* 86:508–514
- Xu H, Dill KA (2005) Water's hydrogen bonds in the hydrophobic effect: a simple model. *J Phys Chem B* 109:23611–23617
- Yang L, Liao R-Z, Yu J-G, Liu R-Z (2009) DFT study on the mechanism of *Escherichia coli* inorganic pyrophosphatase. *J Phys Chem B* 113:6505–6510
- Zaitseva E, Yang ST, Melikov K, Pourmal S, Chernomordik LV (2010) Dengue virus ensures its fusion in late endosomes using compartment-specific lipids. *PLoS Pathog* 6:e1001131
- Zhang JZ (2000) Protein-length distributions for the three domains of life. *Trends Genet* 16:107–109
- Zhang Y, Cremer PS (2006) Interactions between macromolecules and ions: The Hofmeister series. *Curr Opin Chem Biol* 10:658–663
- Zimmerberg J, Kozlov MM (2006) How proteins produce cellular membrane curvature. *Nature Rev Mol Cell Biol* 7:9–19

# Chapter 9

## Orthology: Promises and Challenges



Yannis Nevers, Audrey Defosset, and Odile Lecompte

**Abstract** Orthology is a cornerstone of comparative genomics and has numerous applications in current biology. In this chapter, we first introduce the concepts of orthology and paralogy. We then present the currently available orthology inference methods and the community-led efforts of standardization and benchmarking accompanying these developments. The large panel of available orthology resources is compared in terms of species coverage, access, contextual data and tools proposed to end-users to facilitate the analysis and exploitation of orthology data. We then review the importance of orthology applications, ranging from the study of protein families and information transfer to the comparison of genomes and genotype/phenotype correlations. Finally, we discuss the current challenges in the orthology field, faced with an ever-increasing number of proteomes of particularly heterogeneous quality. We highlight the urgent need of considering orthology at the protein domain and transcript levels and the conceptual and practical difficulties that this raises.

---

Y. Nevers · A. Defosset · O. Lecompte (✉)  
Complex Systems and Translational Bioinformatics, ICube UMR 7357,  
Université de Strasbourg, Strasbourg, France  
e-mail: [odile.lecompte@unistra.fr](mailto:odile.lecompte@unistra.fr)

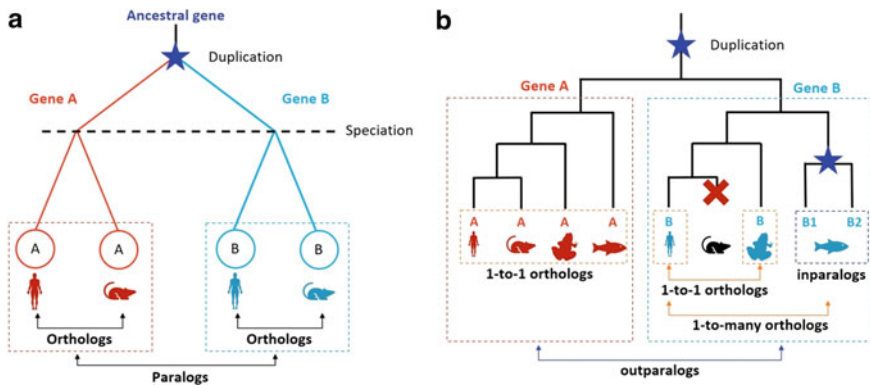
Y. Nevers  
e-mail: [yannis.nevers@unil.ch](mailto:yannis.nevers@unil.ch)

A. Defosset  
e-mail: [adefosset@etu.unistra.fr](mailto:adefosset@etu.unistra.fr)

Y. Nevers  
SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland  
Department of Computational Biology, University of Lausanne, Lausanne, Switzerland  
Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

## 9.1 Introduction

Homology is a central concept in biology and is essential for any intraspecies or interspecies sequence comparison. Originally employed to compare phenotypic traits, it is now mainly used to define relationships between genomic regions, genes and, by extension, between proteins or even sub-protein regions. In this context, homology describes the relationship between two molecular entities (usually genes or proteins) that descend from the same ancestor. Two main categories of homologs were distinguished in the early days of molecular biology (Fitch 1970): paralogs that derive from a common ancestor by a duplication event and orthologs that emerge after a speciation event (Fig. 9.1a). *Stricto sensu*, these definitions only refer to the evolutionary history of genes. However, it is commonly accepted that orthologs tend to retain a similar function, while paralogs may have different fates in the course of evolution. Indeed, the paralogous copies may develop more specialized functions compared to the ancestral gene (tissue/stage-specific expression, complementation of functions initially performed by a single gene) or one copy may evolve a new function under the reduced selection pressure or even degenerate into a pseudogene (Force et al. 1999). The ‘orthology conjecture’ states that orthologs frequently retain ancestral function while paralogs tend to diversify is widely used to transfer functional information between orthologs. Although this hypothesis is commonly accepted by the community, it has been challenged in some cases (Studer and Robinson-Rechavi 2009; Nehr



**Fig. 9.1** Homology relationships. **a** Evolutionary history of a gene family with duplication and speciation events. Genes A (in red) present in humans and mouse emerged after a speciation event, they are orthologous to each other. The same is true for genes B (in blue). Genes A and B are paralogous between each other because they are separated by a duplication event in their evolutionary history. **b** Genes A (in red) are only separated by speciation events, they are 1-to-1 orthologs. The evolutionary history of genes B (in blue) is more complex with a lineage-specific loss in mouse and a ‘recent’ duplication in fish. Considering the evolutionary history of vertebrates, genes B1 and B2 are inparalogs to each other and co-orthologs to the human gene B. Thus, there is a 1-to-many orthology relation between the human gene B and the fish genes B1 and B2 genes. Considering Vertebrates, genes A and B are outparalogs between each other because they emerged after a duplication that occurred in the vertebrate ancestor, i.e., before speciations

et al. 2011), especially among highly similar genes. Nevertheless, it still generally holds (Altenhoff et al. 2012; Chen and Zhang 2012). Notably, it has been shown that the organization of introns (Henricson et al. 2010), the three-dimensional structure of proteins (Peterson et al. 2009) and domain architecture (Forslund et al. 2011) tend to be more conserved between orthologs than paralogs. In addition, orthologs are generally expressed in the same tissues in contrast to paralogs (Kryuchkova-Mostacci and Robinson-Rechavi 2015).

The debate around the orthology conjecture underlines the importance of taking into account the chronology of speciation and duplication events to establish functional links between homologous genes. Indeed, paralogs that derive from a ‘recent’ duplication event may still share the same function in contrast to distant paralogs separated over millions of years of evolution. Unfortunately, there is no objective threshold to define recent versus ancient paralogs, and in fact, it all depends on the evolutionary distance between compared species. This has been conceptualized with the terms ‘outparalogs’ and ‘inparalogs’ coined in 2002 (Sonnhammer and Koonin 2002). When comparing two species, paralogs deriving from a duplication event that occurred prior to the speciation event are called outparalogs, while paralogs originating from a duplication event subsequent to the speciation event are called inparalogs. Inparalogs are considered to be co-orthologs of genes descending from the speciation event in the other species (Fig. 9.1b). Hence, inparalogy and outparalogy are relative notions: The same paralogous sequences can be considered inparalogs or outparalogs depending on the speciation referred to. The co-orthology concept also introduces different orthology relationships: 1-to-1, 1-to-many and many-to-many orthologs (Fig. 9.1b).

The characterization of these intricate homology relationships is far from trivial since there is no direct record of past speciation or duplication events, and evolutionary scenarios can be further complicated by lineage-specific gene losses, whole genome duplications (WGD) and horizontal gene transfers (HGT). WGD or polyploidy can arise within a single species by the doubling of the chromosome set (autopolyploidy) or can result from the merging of the chromosome sets of two different species and subsequent genome doubling (allopolyploidy) (see Van de Peer et al. 2017 for a recent review). Homologs arising by autopolyploidy are called ohnologs (Wolfe 2000) and constitute a special case of paralogs, since both copies evolved originally in the same genomic context. Homeologs that result from an allopolyploidy event are more complex to define (reviewed in Glover et al. 2016) but are observed in many plants. Like orthologs, they originally emerge after a speciation event, but they are subsequently integrated in a single genome through autopolyploidization. Thus, homeologs experience a mosaic fate by initially evolving like orthologs and then after hybridization, undergoing an evolutionary pressure usually exerted on paralogs.

In HGT, the relationship does not rely on vertical transmission of genes but on acquisition of genetic material from another species. Genes whose history since their common ancestor involves an horizontal transfer are called xenologs (Gray and Fitch 1983; Fitch 2000). Xenology is especially prevalent in prokaryotes with HGT frequently involving mobile genetic elements, but it can also occur between



prokaryotes and eukaryotes (notably in the case of endosymbiosis or endoparasitism) or even between eukaryotes (reviewed in Soucy et al. 2015). Xenology relationships encompass a wide range of evolutionary histories, and xenolog classes have been proposed to reflect the events associated with the divergence of xenologs and the relative timing of transfer and speciation events (Darby et al. 2017).

The first step in the process of characterization of homology relations is based on sequence comparison. It is assumed that genes/proteins are homologous if they exhibit a higher similarity than would be expected by chance. Thus, homology detection usually relies on similarity searches, typically a BLAST search (Altschul et al. 1997; Camacho et al. 2009), with a fixed threshold of score, percentage identity, expect-value, etc. The distinction at the genome scale between the different types of homology (1-to-1 orthology, co-orthology, inparalogy, outparalogy, xenology) then requires dedicated approaches. The methods used to infer orthology and the corresponding available resources are presented in the first section of this chapter. We then review the main applications of orthology in biology. In the last section, we highlight the practical and conceptual challenges around the notion of orthology and its uses.

## 9.2 Orthology Inference and Resources

### 9.2.1 Orthology Inference Methods

An exhaustive description of the plethora of available programs is beyond the scope of this review (for a recent review on methods, see Altenhoff et al. 2019). However, these different programs can be classified into four main categories: graph-based, tree-based, hybrid and meta-prediction methods that are presented briefly below.

In graph-based methods, genes/proteins are represented by nodes and homology relationships by edges in the graph. The graph construction relies on all-against-all similarity searches between genes/proteins from two genomes. The simplest approach, called reciprocal best hit (RBH), will predict an orthology relationship between proteins A and B from two genomes if A is the genome-wide closest relative of B and vice versa (Overbeek et al. 1999). This approach only considers 1-to-1 orthology relationships, thus overlooking one-to-many and many-to-many orthologs. To circumvent this problem and offer a more comprehensive view of evolutionary relationships, other algorithms have been developed where inparalogy relations are inferred and included during graph construction. Examples of such methods include COG (Tatusov et al. 1997), Inparanoid (Remm et al. 2001), OrthoMCL (Li et al. 2003), OMA (Roth et al. 2008), EggNOG (Jensen et al. 2008), OrthoInspector (Linard et al. 2011) and OrthoFinder (Emms and Kelly 2015). The homology relationships predicted between a pair of genomes can then be extended to a set of species, in order to define groups of orthologs (also called orthogroups) present in these species. The groups are delineated on the basis of the structure of the graph by transitivity or clustering. For instance, OrthoMCL uses Markov clustering to partition the homology

graph into orthogroups containing highly connected orthologs and recent paralogs. OMA groups are based on cliques, i.e., fully connected subgraphs corresponding to genes that are all orthologs to each other, thus de facto excluding orthologs involved in 1-to-many or many-to-many relations.

Tree-based methods infer orthologs based on the gene's evolutionary history, which is reconstructed by reconciling the gene family tree with the species tree. First, a multiple alignment of homologous sequences is constructed to generate a phylogenetic tree of the gene family. Then, the nodes of this gene tree are labeled as duplication or speciation events by comparison to the species tree during the reconciliation step, allowing the prediction of orthology and paralogy relationships. This type of approach is implemented in numerous programs, including RIO (Zmasek and Eddy 2002), Orthostrapper (Storm and Sonnhammer 2002), PhylomeDB (Huerta-Cepas et al. 2007), Ensembl Compara (Vilella et al. 2009), PANTHER (Mi et al. 2010). These methods produce hierarchical ortholog groups, i.e., groups of orthologs and inparalogs deriving from a common ancestor, in the form of trees. These hierarchical groups are more informative than simple orthology relationships between pairs of species or flat groups of orthologs without evolutionary information about intra-group relations. Unfortunately, tree-based methods are highly dependent on the construction of correct multiple alignments and trees and are computationally demanding, preventing their application to very large datasets.

Although hierarchical groups are naturally produced by tree-based methods, they can also be generated by a post-processing of orthogroups obtained by graph-based methods. As an example, EggNog and OrthoDB explicitly delineate the hierarchy of ortholog groups by identifying orthogroups at different taxonomic levels of the species tree. Hybrid methods go further by using attributes of graph-based and tree-based methods in the inference of orthology relationships itself. The method of OMA Hierarchical Orthologous Groups (HOG) (Altenhoff et al. 2013) uses an orthology graph of pairwise relations to form groups, starting with the most specific taxonomic level and progressively merging groups toward the root of the species tree. Hieranoid (Schreiber and Sonnhammer 2013) progressively calculates pairwise orthology relationships using RBH at each node of a guide tree from the leaves to the ancestor. At each node, a consensus or a profile is built from the child nodes and used for subsequent pairwise comparisons, which considerably reduces the number of required pairwise comparisons. OrthoFinder 2 (Emms and Kelly 2019) first identifies orthogroups among a set of species using the OrthoFinder graph-based approach (Emms and Kelly 2015) and then uses the orthogroups to infer approximate gene trees and a species tree. Finally, each gene tree is compared to the species tree to infer duplication events and refine prediction of orthology and paralogy relations.

Meta-prediction methods are designed to exploit predictions generated by different programs and thus can potentially highlight false positives and negatives. As an example, DIOPT (Hu et al. 2011) assigns a score to each orthology relationship according to the number of independent methods predicting this relation. The MARIO program (Pereira et al. 2014) goes further by delineating a group of orthologs from predictions of several methods and constructing a hidden Markov model (HMM) profile of these orthologous sequences. This profile is then used

to evaluate the predictions made by each individual method. MetaPhOrs (Pryszcz et al. 2011) integrates phylogenetic trees constructed by several methods to predict orthology relations and assigns a score depending on the number of predictions. This filters unreliable results linked to poor resolution of phylogenetic trees. The WORMHOLE program (Sutphin et al. 2016) uses a classifier based on support vector machines (SVM) trained on a positive set of validated orthology relationships and a negative set of non-orthology gene pairs. The algorithm assigns a weight to each prediction method depending on its performance in different test cases (e.g., according to the proximity of the species under consideration). This weight is then used to combine predictions on a complete dataset and extract reliable orthology relations.

### 9.2.2 *Standardization and Benchmarking*

Given the multiplicity of orthology inference methods available, it is crucial to cross-reference, compare and evaluate their predictions in different biological contexts in order to choose the relevant program for a given biological question and to improve prediction methods. This requires a standardization of orthology prediction formats and an objective benchmarking. These topics are the central goals of the Quest For Orthologs (QFO) consortium (Gabaldón et al. 2009). QFO addresses both conceptual issues and technical challenges in orthology prediction. For example, community efforts led to the development of the standardized OrthoXML format (Schmitt et al. 2011) designed to represent orthology predictions for both graph- and tree-based methods. An ontology (Fernández-Breis et al. 2016) has also been developed to formalize the representation of orthology relationships. This ontology allows the representation of data according to a semantic Web standard, resource descriptions framework (RDF) that facilitates interoperability between resources.

The QFO consortium has also defined a QFO reference proteome dataset to allow the comparison of methods on a common set of species and proteins. The dataset is updated every year and currently comprises 78 UniProt Reference proteomes. It includes sequences from model organisms, species of interest for biomedical or agronomic research or species of interest from a phylogenetic point of view (Sonnhammer et al. 2014). In parallel, a variety of benchmarks have been developed to evaluate orthology prediction methods according to phylogenetic and functional criteria. A large-scale benchmarking study (Altenhoff et al. 2016) comparing 15 orthology methods highlighted a trade-off between sensitivity and specificity and clearly showed that the best approach is highly dependent on the biological context. Overall, the orthogroup predictions of OMA are characterized by high specificity, whereas the tree-based method used in PANTHER has high sensitivity. However, there is no systematic difference between tree-based and graph-based methods. Finally, Inparanoid, Hieranoid and OrthoInspector as well as OrthoFinder in the most recent version of the benchmark (results available at <https://orthology.benchmarkser>

[vice.org](https://www.vice.org)) show a good balance between specificity and sensitivity over all benchmarks. Orthology predictions from the best methods identified by the benchmarking are now integrated in the Alliance of Genome Resources (Alliance) portal (Alliance of Genome Resources Consortium 2020). The Alliance aims to facilitate exploration of orthologous genes in human and well-studied model organisms in order to exploit the wealth of genetic and genomic studies available in these organisms.

### 9.2.3 Orthology Resources

Most orthology inference programs can be installed and executed locally on a user-defined set of proteomes, but many of them are also used to generate databases of orthology relationships. These resources are essential for the routine use of the orthology concept by non-experts. The databases differ in terms of number and diversity of represented species (Table 9.1), which determines the granularity with which orthology relationships can be exploited. Some generalist databases cover a large panel of species such as EggNog (Huerta-Cepas et al. 2016), HOGENOM (Penel et al. 2009), Inparanoid (Sonnhammer and Östlund 2015), MBGD (Uchiyama et al. 2019), OMA (Altenhoff et al. 2018), OrthoDb (Kriventseva et al. 2019) and OrthoInspector (Nevers et al. 2019). EggNog and OrthoDB also include viral genomes. Other resources are clade-specific, including TreeFam (for Metazoa) (Schreiber et al. 2014), FungiPath (for Fungi) (Grossetête et al. 2010), and GreenPhylDB (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018) that focus on plants. With the exception of MetaPhOrs (Pryszcz et al. 2011), the resources based on meta-predictions generally focus on a small number of model species (Table 9.1). In addition to the databases dedicated to orthology, orthology relationships are also provided in more general biological portals, such as PANTHER (Mi et al. 2019), Ensembl Compara (Herrero et al. 2016) and HomoloGene (NCBI Resource Coordinators 2016).

Orthology databases offer diverse access to information, via Web interfaces for manual exploration or using programmatic access through Web services or SPARQL (SPARQL Protocol and RDF Query Language) interfaces. Users can search for orthologs of a given gene using genes/proteins or orthogroup identifiers or perform a sequence similarity search. Information can also be accessed through functional annotation of the gene of interest (keywords, description or GO annotations). For instance, OrthoInspector (Nevers et al. 2019) allows users to retrieve all proteins of a given species associated with a given GO term and visualize their evolutionary histories. OrthoMCL (Chen et al. 2006) and GreenPhylDB (Rouard et al. 2011) propose searches for groups with a given protein domain. Genes can also be retrieved on the basis of their phylogenetic distribution, i.e., the presence or absence of an ortholog in different taxa. This phylogenetic profiling search is implemented in MBGD (Uchiyama et al. 2019), OrthoDb (Kriventseva et al. 2019), OrthoInspector (Nevers et al. 2019), OrtholugeDB (Whiteside et al. 2013), OrthoMCL (Chen et al. 2006) and GreenPhylDB (Rouard et al. 2011). It can be used to perform genotype/phenotype studies as discussed in the applications section.

**Table 9.1** Main orthology resources

Resource		Coverage					Exploration						Representation								
Type	Name	Genomes	Bacteria	Eukaryota	Archaea	Viruses	Gene Id	Group Id	Sequence	Function	Distribution	SPARQL	Webservice	Orthologues	Function	Domains	MSA	Tree	Synteny	Distribution	
General	Inparanoid	273	/	/	/	0	✓	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
	OMA	2 327	1688	485	154	0	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓
	EggNOG	2 031	1678	115	238	352	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓
	OrthoDb	7284	5609	1271	404	7963	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
	OrthoMCL	150	36	98	16	0	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗
	Hieranoid	66	20	40	6	0	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
	OrthoInspector	4753	3863	711	179	0	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✓
	MBGD	6318	5861	203	254	0	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓
	OtholugeDb	2069	/	0	/	0	✓	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗
HOGENOM	13367	12326	593	224	0	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	
PhylomeDb	1 862	/	/	/	0	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓	✓	✗	
Specific	TreeFam	109	0	109	0	0	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	
	FungiPath	165	0	165	0	0	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✗	✗	
	Greenphyl	37	0	37	0	0	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	
	PLAZA	119	0	119	0	0	✓	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	
Meta-predictions	P-POD	12	1	11	0	0	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓	
	MetaPhOrs	2713	1 720	877	116	1	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	
	WORMHOLE	6	0	6	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	
	DIOPT	10	0	10	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	
	YOGY	11	1	10	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	
HCOP	19	0	19	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗		
Other	Panther	142	35	99	8	0	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓	✗	✗	
	Ensembl	1191	123*	1068	/	0	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	
	Homologene	21	0	21	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓	

References: EggNog (Huerta-Cepas et al. 2016), HOGENOM (Penel et al. 2009), Inparanoid (Sonnhammer and Östlund 2015), MGD (Uchiyama et al. 2019), OMA (Altenhoff et al. 2018), OrthoDb (Kriventseva et al. 2019), OrthoInspector (Nevers et al. 2019), OrtholugeDB (Whiteside et al. 2013), OrthoMCL (Chen et al. 2006), PhylomeDB (Huerta-Cepas et al. 2014), TreeFam (Schreiber et al. 2014), FungiPath (Grossetête et al. 2010), GreenPhylDB (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018), PANTHER(Mi et al. 2019), Ensembl Compara (Herrero et al. 2016), HomoloGene (NCBI Resource Coordinators 2016)

\*123 prokaryotic species (mainly Bacteria but also some Archaea) are included in the Pan-Compara resource which includes a selection of prokaryotic and eukaryotic species

All orthology databases provide orthology predictions in the form of a list of orthologs in the covered species, but many of them contextualize this minimum information by adding relevant data and tools to analyze and exploit the evolutionary information (Table 9.1). Hence, they frequently provide additional information about the function (GO term annotation, enzyme classification numbers...) or architecture (protein domains) of the predicted orthologs as illustrated in Table 9.1. This functional information most often comes from automatic annotations that must be handled with care. However, viewing the annotations for all the orthologs of a protein makes it easier to detect inconsistencies and spurious annotations. For example, OMA (Altenhoff et al. 2018) offers a synthetic representation of the GO annotations

of the detected orthologs with a color code that distinguishes between automatic annotation, annotation validated by an expert and annotation based on experimental data. Multiple sequence alignment (MSA) and phylogenetic trees also constitute an essential analytical tool for a more in-depth understanding of the relationships between orthologs and paralogs. As such, they are often made available, in particular by tree-based methods. They are either pre-calculated and available directly on the Web interface or can be constructed ‘on the fly’ for a selection of predicted orthologous sequences. In addition, some resources provide information about the genomic context of the query gene and its orthologs, allowing to detect syntenic stretches of genes that can be helpful for the validation of orthology relations and may be indicative of a functional link between syntenic genes. Finally, orthology resources can provide the taxonomic distribution of detected orthologs in each species represented in the orthology database. This is suitable for clade-specific resources such as GreenPhylDb (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018). For generalist orthology resources, a synthetic view of distributions is required as exemplified by OrthoInspector (Nevers et al. 2019) that provides schematic representations of phylogenetic distributions at different granularity levels.

## 9.3 Orthology: The Swiss Army Knife of Genomics

### 9.3.1 *Exploration of Gene and Protein Families*

Since their definition in the early seventies, orthologs and paralogs have been traditionally used to study gene and protein families, in particular in the framework of multiple alignment analysis. By placing a gene or a protein sequence in its evolutionary context, the multiple alignment reveals selection pressure existing at particular sequence positions, allowing the straightforward detection of conserved motifs, localization signals or key functional residues for a considered family of orthologs or a superfamily regrouping several paralogous families (Lecompte et al. 2001). Such evolutionary analyses are essential for the determination of catalytic sites or residues involved in protein interactions for example. This can be exploited to decipher residues, motifs or domains involved in the specificity of paralogous families, for instance, to identify residues responsible for the enzyme substrate specificity in a multienzymatic family. In addition, alignments of orthologs or homologs are exploited in both 2D and 3D structure prediction methods by comparative protein modeling (reviewed in Khan et al. 2016). With the increase of experimentally determined structures, a wide range of accurate models are now available that can be used to predict protein binding sites, effects of protein mutations, and for structure-guided virtual screening (Liu et al. 2011; Leelananda and Lindert 2016).

Orthologous sequences are directly exploited by many mutation analysis tools, such as PolyPhen (Adzhubei et al. 2010) or SIFT (Vaser et al. 2016), to predict the phenotypic effects of variants. Pairwise or multiple alignments of orthologous

sequences are also used at the genomic level to highlight conserved regions that may reflect the existence of functional elements. Orthologs are also the cornerstone of phylogenetic studies aimed at deciphering the evolutionary history of a gene family or, more generally, phylogenetic relationships between species. The reconstruction of phylogenetic relationships between species has for a long time relied on a single family of genes, typically 16S/18S rRNA genes or well-conserved housekeeping protein genes. Today, species phylogenies can be built using comparisons of several protein families, including genome-wide comparisons (Crawford et al. 2012). These studies generally focus on widely conserved protein families exhibiting one-to-one orthology relationships. Orthofinder directly exploits orthogroups within a species set to construct a phylogenomics species tree using the species tree from all genes (STAG) algorithm (Emms and Kelly 2018). With the multiplication of available genomes and metagenomes, such phylogenomics analyses have renewed our vision of the tree of life, for instance, by highlighting the bacterial diversification (Hug et al. 2016), reshaping the eukaryotic tree (Burki et al. 2020) and revealing a new group of Archaea, the Asgard that questioned the position of Eukaryotes in the tree of life (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017).

Orthologous sequences and phylogenetic trees can also be exploited for ancestral sequence reconstruction. The leaves of the phylogenetic tree represent the extant sequences of the family, while the root corresponds to the extinct common ancestor. The ancestor can be synthesized to experimentally explore its biochemical properties. This approach allows to resurrect an ancestral precursor with selected properties, such as thermostability, in order to initiate synthetic evolution experiments (Gumulya and Gillam 2017). It can also be used to decipher past environmental conditions. For example, the reconstruction of translation elongation factors from organisms that lived 3.5 billion years ago revealed that the thermostability of these factors declines in the course of evolution and suggested a 30 °C decrease in environmental temperature (Gaucher et al. 2008). Ancestral sequence reconstruction methods also deduce the sequences present at each internal node of the tree. These intermediate states can help to elucidate evolutionary processes, in particular the main mutations involved in the distinct properties of extant proteins (Straub and Merkl 2019). Applied to whole genomes, ancestral reconstruction offers a partial view of ancestral gene repertoires, from the known repertoires of extant species. Such a resource is available on the ancestral genome portal, constructed from PANTHER inferences (Huang et al. 2019).

### 9.3.2 *Information Transfer*

As stated above, orthologous genes tend to retain equivalent functions across species and are thus widely used to transfer information from model species to poorly characterized ones. Typically, the functional annotation of genes in a newly sequenced genome is carried out by identifying annotated orthologs using similarity searches in protein databases such as UniProt (The UniProt Consortium 2019) or through the

Gene Ontology (The Gene Ontology Consortium 2019) and then transferring these annotations to genes of unknown function. Several protocols (compared in Amar et al. 2014) allow this automated annotation transfer. Although this approach is time efficient, it can also lead to bias since the orthology conjecture is not an absolute law and the ortholog/paralog distinction is not trivial, especially in superfamilies (Schnoes et al. 2009). The problem of misannotation is also particularly severe, with multi-domain proteins exhibiting a differential conservation of some domains (discussed in Sect. 9.3.3 Beyond gene level orthology). In addition, automated transfer can propagate annotation errors. It is therefore wise to rely on closely related orthologs with expert-curated annotations, whenever possible, to avoid the ‘percolation of annotation errors’ modeled by Gilks and colleagues and its deleterious effects on database quality (Gilks et al. 2002).

More generally, orthology can be used to transfer experimentally evidenced information obtained from one species to another, provided that the organisms are sufficiently close. This principle is used by the Gene Ontology Consortium (Ashburner et al. 2000; The Gene Ontology Consortium 2019) to propagate standardized annotations not only on protein molecular function but also on their sub-cellular localization and the biological processes in which they are involved. The resulting annotations receive the IEA evidence code (Inferred from Electronic Annotation) in the case of an automatic transfer between orthologs. The Gene Ontology also integrates a semi-automated transfer protocol (Gaudet et al. 2011), taking into account annotations from several orthologs and the phylogenetic relationships between the corresponding species. These annotations are labeled with the IBA code (Inferred from Biological ancestry).

Information about protein–protein interactions (PPIs) can also be transferred from one species to another through the concept of interologs. The term ‘interolog’ (Walhout et al. 2000) refers to the conserved interaction between two pairs of proteins A1 and B1 from a first species and A2 and B2 from a second species. The A1/B1 interaction is considered as an interolog of the A2/B2 interaction if A1 and A2 are orthologs to B1 and B2, respectively. The concept of interology can be exploited in a predictive way: Orthologs of interacting proteins in one species are identified, and the PPI information is transferred to the pair of orthologs. To avoid false positive errors, interology inferences are usually combined with other data, as illustrated by the STRING interaction database (Szklarczyk et al. 2019) that relies on a large panel of diverse evidence (experiments, text mining, co-expression, synteny, etc.).

Finally, when working on human genes, orthology relationships are key elements to consider when choosing a relevant model species for experimental studies. In addition to practical considerations (duration, cost, etc.), the model species should be chosen to avoid 1-to-many or many-to-1 orthology relations between the human and the model species, since the existence of additional inparalogs in one species would considerably complicate the interpretation of experimental results.



### 9.3.3 Comparison of Genomes and Proteomes

Comparisons of complete genomes and proteomes are intrinsically linked to the proper delineation of orthologs and paralogs. Comparisons of orthologs at the sequence level are used to evaluate the selection pressure acting to model evolutionary rates in different species. One of the first examples of such genome-wide assessment of evolution rates was carried out in mammalian and nematode lineages (Castillo-Davis et al. 2004). This study showed that strong purifying selection seems to act on the same central cellular processes (such as translation and protein transport) in mammals and nematodes, whereas positive selection acts on different biological processes in each lineage (DNA-dependent transcriptional regulation in nematodes, signal transduction via receptors and host immune response in mammals). Such comparative analyses are also performed for non-coding RNA genes such as microRNA. For example, the study of microRNA substitution rates in human and chimpanzee genomes revealed that primate-specific microRNAs have twice as many substitutions as older microRNA families (Santpere et al. 2016).

Comparison of proteomes in terms of gene content has become a quasi-obligatory step when sequencing a new genome. It requires the prediction of orthology and paralogy relations between the proteomes under consideration and reveals the set of conserved protein families but also the acquisitions and losses that have taken place independently in each lineage. These comparisons have highlighted the extraordinary plasticity of the gene repertoire among species. This is particularly striking in the case of prokaryotic genomes. In a comparison of more than 500 bacterial species, Lapierre and Gogarten (2009) showed that the conserved bacterial core was reduced to about 250 gene families, with the notable exception of certain symbionts exhibiting a particularly reduced genome. This diversity of gene repertoire observed even among closely related species can be explained by lineage-specific expansion of gene families, acquisition of genes by horizontal transfer (xenologs) and differential gene losses. In some prokaryotes, the genomic versatility is so important that large differences in gene content can occur between different strains of the same species. This led to the definition of the pangenome concept, i.e., the set of all genes present in a given species, that can be divided into the conserved core and the accessory genome (reviewed in Brockhurst et al. 2019). In species with an ‘open’ pangenome, the core genome conserved in all strains represents only a small fraction of the pangenome, questioning the concept of species in Prokaryotes. For instance in *Escherichia coli*, the core genome is restricted to about 3000 gene families, while the pangenome reaches a total of about 90,000 families (Land et al. 2015).

Comparisons of orthologous genomic regions or complete chromosomes decipher the evolution of genome architecture by revealing differential gains/losses of genomic regions, segmental duplications and balanced rearrangements. These comparisons can be made at the nucleotide level using, for example, BLASTZ (Schwartz et al. 2003) or LASTZ and chaining/netting programs (Kent et al. 2003) to discriminate between orthologous and paralogous alignments. Alternatively, the comparison of genomic regions can be based on the comparison of genomic location of orthologs in

different genomes to identify conserved syntenic blocks, i.e., a stretch of genes with a conserved gene order in different species. Such comparisons delineate syntenic genes frequently linked by functional relations and allow the detection of elements involved in genomic plasticity at the syntenic regions boundaries. They are also used to reconstruct ancestral genomes with distance/event-based or homology/adjacency-based methods (reviewed in Feng et al. 2017).

### 9.3.4 *Functional Inferences and Genotype/Phenotype Correlations*

Comparisons of complete proteomes based on orthology relationships can be exploited to perform functional inferences between genes or to detect genes potentially involved in a phenotype. The rationale behind this approach is that functionally linked genes are preserved or lost in a correlated manner over the course of evolution and thus are found in the same species (Pellegrini et al. 1999). This assumption can be exploited in different ways. Subtractive analysis aims to identify genes restricted to species with a given phenotype. In practice, this means comparing the gene repertoire of at least two species (species A and B) possessing the phenotypic trait of interest and one or several related species (species C) lacking the considered phenotype. The set of genes with orthologs in species A and B but without orthologs in species C is likely to be enriched in genes associated with the phenotypic trait of interest. This approach was introduced by Huynen (Huynen et al. 1998) in the early days of comparative genomics in order to compare the genome of the pathogen *Helicobacter pylori* with that of another pathogen *Haemophilus influenzae* and a benign strain of *E. coli*. They identified 17 gene families restricted to the pathogenic species and potentially involved in virulence and host–pathogen interactions.

The subtractive method is applicable to the search for genes linked to a phenotypic trait or biological process that has been lost/acquired in some species during evolution. This approach can be extended to the comparison of tens or hundreds of genomes to allow a precise definition of the phenotypic distribution. The comparison of phylogenetically distinct lineages that have independently acquired (or lost) a given phenotype limits false positive predictions by eliminating genome differences simply due to random gains and losses of genes. For instance, Hecker and colleagues (Hecker et al. 2019) compared mammalian genomes to identify convergent gene losses associated with dietary adaptations in six independent herbivore lineages (16 species) and five independent carnivore lineages (15 species). Regarding the small evolutionary distances separating these placental mammals, they considered not only loss of entire genes or exons but also gene-inactivating mutations, using a genomic approach that combines the identification of orthologous regions and the CESAR program, a coding exon-structure aware realigner (Sharma et al. 2016).

At a larger evolutionary scale, another methodological framework is required. Phylogenetic profiles represent a generalization of subtractive analysis allowing the comparison of a large number of genomes that can be evolutionary distant. A phylogenetic profile of a gene represents the presence or absence of orthologs of that gene in the genomes of several species (Tatusov et al. 1997). Phylogenetic profiles were first used to infer the function of uncharacterized genes, and the method has been successfully applied to the annotation of genes, mainly prokaryotes (see Kensche et al. 2008 for examples). They are also exploited to predict functional links between genes, notably in the STRING (Szklarczyk et al. 2019) and OrthoInspector databases (Nevers et al. 2019).

Phylogenetic profiles can not only be compared to each other but also to all types of presence–absence distributions, including phenotypic traits. Phylogenetic profiling can thus be exploited to perform phenotype-genotype association studies. One of the first studies of this type was carried out on 86 prokaryotic genomes to identify genes associated with thermophily (Jim et al. 2004). Since then, many similar studies have been performed, notably to identify genes involved in human diseases thanks to the huge increase of available eukaryotic genomes that allows a detailed exploration of the distribution of human genes. For instance, Tabach et al. (2013) identified 54 clusters of phylogenetic profiles associated with a specific class of symptoms. More recently, the profiling of human genes in 100 eukaryotic species revealed 274 human genes exhibiting a phylogenetic distribution correlated with the distribution of cilia in eukaryotic lineages (Nevers et al. 2017). This set of predicted ciliary genes includes 87 new candidates. Among them, 21 have already been experimentally validated as ciliary genes.

## 9.4 Challenges

### 9.4.1 *Keeping Up with the Data Flow*

As seen above, orthology is the cornerstone of a plethora of applications in comparative genomics and biology, and orthology resources provide numerous contextual data and analytical tools to facilitate orthology exploitation. Coming into a new decade, they are now gearing up to adapt to new challenges, a data flow brought by the next generation sequencing and a need to assess orthology at different granularity levels. The last two decades have seen a massive increase in sequencing capacities, leading to the acquisition of numerous genomes from across the tree of life. These genomes have obvious usefulness for studying evolution at a broad scale and are increasingly incorporated into orthology resources. Nevertheless, they also lead to important challenges linked to the management and analysis of the ever-increasing volume of data and the heterogeneous data quality.

Genomic data, hence genome annotations, have been increasing at an exponential rate with the advent of high-throughput sequencing technologies. As of today, 19,163 complete genomes are registered in the Genome Online Database (Mukherjee et al. 2019), as well as 215,613 genomes in the permanent draft state. This increase in data generation represents a challenge for orthology resources. It is especially true for tree-based approaches, which are commonly more computationally intensive as they rely on phylogenetic tree inference tools and are traditionally limited in the number of species they can include. While less computationally intensive, the data increase is still onerous for graph-based approaches, as they rely on all-vs-all sequence comparisons, which grow quadratically with the number of sequences. The legacy tools for these kinds of comparison, namely BLAST (Altschul et al. 1990; Camacho et al. 2009) or Smith-Waterman (Smith and Waterman 1981) alignment, do not scale well, and resources that use them rely heavily on high-performance computing clusters. Other tools and resources use faster but generally less sensitive solutions: MMSeq2 (standard modes) (Steinegger and Söding 2017), DIAMOND (Buchfink et al. 2015) or *ad-hoc* methods as in SwiftOrtho (Hu and Friedberg 2019) for instance can perform all-vs-all comparisons with better performances.

Nonetheless, solutions bypassing computationally intensive all-vs-all computations are increasingly being investigated, in anticipation of an even bigger surge in data. These approaches such as EggNog-Mapper (Huerta-Cepas et al. 2017) aim to reduce the computation required to adding new proteomes by exploiting already precomputed ortholog groups that are assumed to be stable over time. Their goal is to use fast methods, e.g., hidden Markov models (Eddy 2011) or k-mer based sequence similarity searches, to identify likely existing orthologous groups in which each sequence fits. While fast, these methods rely on existing databases with sufficient clade coverage to be efficient.

Another aspect of data management, linked to computational time, is the size of databases produced. Storing a high number of orthologous relations or orthologous size implies storing Terabytes of data and induces longer access times to the data. Consequently, it is not necessarily optimal for orthology resources to include all available genomes, and a choice is often made concerning which data to select, with high variability of species chosen in each orthology resource. This is reflected by the number of species available in different resources and variable representation in terms of clades or domains of life. Notably, some resources specialize in specific clades such as Plaza (Van Bel et al. 2018) for plants or FungiPath for fungi (Grossetête et al. 2010). Even among the databases with a large number of species, a wide diversity of species is preferred rather than sheer number, as diversity is generally more important than number in comparative studies (Škunca and Dessimoz 2015). This can be achieved by limiting additional species to new taxa of interest or by limiting inter-clade computations to fewer species (Nevers et al. 2019) with several levels of taxonomic resolution. The decision to add or keep a species in an existing database is a product of multiple factors but may be informed by indicators of how the addition of one species affects the diversity. For example, the rarefaction curve proposed by the KinFin analysis tool (Laetsch and Blaxter 2017) (compatible with some orthology inference software suites) provides an objective measure of the novelties in terms

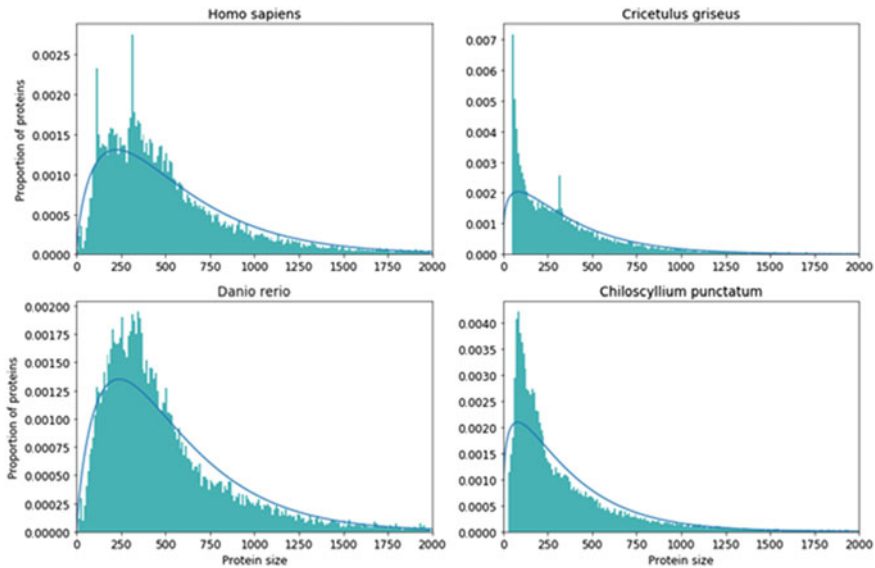
of orthogroups added by each included species. Favoring diversity is also beneficial for the fast-placing strategies mentioned above, moving toward resources with a limited number of species computed directly with the all-versus-all strategy, and other species added to existing groups using less computationally intensive strategies.

### 9.4.2 Addressing Proteome Quality

Another aspect of high-throughput data is the associated data quality issues, and the genomic data used in comparative genomics studies are no exception. Proteomes, i.e., the genome annotations of protein coding genes from genomic data result from a multi-step process ranging from genome sequencing to the actual annotation of the final assembled sequences, with multiple possible sources of error. Consequently, a proteome may have missing proteins (either being permanent draft or a misannotated complete genome) or contain proteins that are either fragmented or actually erroneous. All of these cases may in turn induce errors in orthology inference that rely heavily on sequence comparison and in comparative genomics approaches that assume data completeness.

Missing proteins, for instance, lead to missing orthology relationships between species with incomplete genomes and other species. Most orthology pipelines assume data completeness when inferring orthology, and while they are in principle robust to gene losses, incomplete gene sets may lead to errors in orthology inference and in orthogroup reconstruction. Some methods, e.g., Hamstr (Ebersberger et al. 2009) and OrthoGraph (Petersen et al. 2017) are designed to avoid this assumption by first excluding incomplete datasets (e.g., issued from RNA-seq data) during orthogroup construction. Sequences from the incomplete datasets are then mapped to the precomputed robust orthogroups. Even with correct orthology inference, incomplete genomes impact the phylogenetic placement of species, as fewer marker genes are available. This is particularly detrimental when relations between species are hard to resolve. More spectacularly, artificially missing proteins constitute a significant source of errors for comparative genomics methods relying on comparison of entire species gene repertoires, e.g., phylogenetic profiling.

Fragmented proteins are another matter and initially have an impact on orthology prediction via sequence similarity comparisons. For example, if a fragmented protein sequence corresponds to a single domain, reciprocal best hit methods may infer a false positive pairwise relation with a protein in another species having a homologous domain, although the full-length protein would not be identified as orthologous. Conversely, if the protein fragment corresponds to a low complexity, repeat-containing or divergent region, similarity based orthology prediction methods will miss it, leading to false negatives and in the worst case, may even be responsible for spurious relations (false positives). It is worth noting that issues caused by this kind of region, amplified in the presence of fragments, constitute a general limit of similarity-based orthology inference methods in any organism.



**Fig. 9.2** Protein length distribution in four proteomes, from various vertebrate clades. On the left are examples of the distribution observed in well-studied species (*Homo sapiens* and *Danio rerio*), similar to the one observed in most proteomes. On the right, examples of atypical distributions with high number of small proteins for the rodent *Cricetulus griseus* and the chondrichthyes *Chiloscylidium punctatum*

A stark difference in proteome data quality is revealed by analysis of the distribution of protein length between publicly available proteomes. For example, Fig. 9.2 shows the protein length distribution, normalized for proteome size, in four vertebrate species. Most proteomes share a distribution centered on a peak in the range of 200–400 amino acids and a decreasing number of long proteins, as illustrated by *Homo sapiens* and *Danio rerio* (Fig. 9.2). In contrast, some proteomes present a peak for small proteins (less than 100 amino acid long), as exemplified by the other proteomes presented on Fig. 9.2. Strikingly, all manually curated proteomes of model species have the former distribution, and both distributions are distributed across the species tree, ruling out biological exceptions (Nevers et al. in prep). Instead, it indicates a high number of truncated or erroneous proteins.

One must thus be cautious when providing annotations of genomic data to public databases or using these data for orthology inference and comparative genomics. Quality measures exist to indicate the quality of genome assembly, N50 being a standard indicator of genome contiguity that is commonly provided with published genome assemblies. However, genome assembly quality does not necessarily correlate with proteome annotation quality. State-of-the-art tools exist that provide an indication of data completeness and fragmentation. For instance, CEGMA (Parra et al. 2007, 2009) and its successor, BUSCO (Waterhouse et al. 2018) make use of known conserved gene families, so-called core orthologs, in single-copy in most species for

the latter, to assess the completeness of the gene annotation for a given genome. The assumption being that the proportion of core orthologs found in a genome reflects the completeness of the gene annotation as a whole. BUSCO provides additional information about the state of the proteome, by indicating which proportion of core orthologs are found only in a fragmented state. Assessing BUSCO completeness is standard practice when publishing new genomes, and this information is now available in UniProt (The UniProt Consortium 2019) for most available proteomes.

However, empirical data show that BUSCO completeness assessment is not always correlated to the standard protein length distribution, suggesting that it does not capture all cases of genome misannotation. A better proxy of this bias can be obtained in the form of summary statistics, such as the proportion of extremely short proteins in the genome or the number of proteins annotated as not starting with a methionine (i.e., annotated genes for which no start codon was found by the annotation pipeline). These summary statistics can be used to filter genomes used in orthology analysis (Nevers et al. 2019), by setting thresholds under which proteomes are considered as not annotated. As these parameters are nearly orthogonal to core ortholog completeness, they can be used in parallel with methods like BUSCO and CEGMA to identify low quality proteomes. Despite these developments, work is still needed to further assess proteome quality and its impact on downstream applications, and this issue is an important target for future community efforts.

### 9.4.3 *Beyond Gene-Level Orthology*

While most orthology prediction methods are based on full-length gene or protein sequences, in certain cases, functional domains might be a more pertinent entity to consider. Indeed, the majority of known proteins consist of multiple domains, especially in the eukaryotic lineages, and it is known that multi-domain architectures tend to evolve over time as a result of different mechanisms, such as domain gains, losses and duplications, or gene fusion and fission (Buljan and Bateman 2009). The latter in particular can result in complex evolutionary histories for genes with domains of very different ancestral origins, which in turn makes orthology relations more complicated. In addition, domain architecture rearrangements have been observed several times between orthologs of species belonging to different phyla, possibly as a consequence of different organism complexity (Koonin et al. 2000, 2004). However, studies have shown that domain rearrangements can occur between relatively close species, such as mammals or members of the *Drosophila* genus, and it has been estimated that they could concern up to 50% of proteins (Forslund et al. 2011; Wu et al. 2012; Sonnhammer et al. 2014).

Divergences of domain content and/or order between orthologs can be challenging for traditional orthology inference methods. In some cases, parts of the protein sequence might be too highly divergent in some species to be properly detected as orthologs. In other cases, one protein might have significant similarity to multiple different protein families, each due to a different domain of the query protein, making

it hard to clearly establish orthologous relations. This shows a clear limitation of full-length analyses, as they ignore the natural tendency of proteins to be modular and to evolve not at the complete sequence level, but at the domain level. It would be beneficial to focus future improvements and developments on domain-aware orthology inference as a complement to full-length methods, in order to predict more precise ortholog relations and better understand architectural rearrangements in protein evolution. While it has been widely acknowledged that such methods are needed (Sjolander et al. 2011), very few currently take domains into account. Exceptions include the microbial genome database MGD, which constructs ortholog groups at the domain level (Uchiyama et al. 2019), and Domainoid (Persson et al. 2019), a tool that uses Pfam (El-Gebali et al. 2019) defined domains to infer orthology relations at the single level domain. Domainoid has been shown to retrieve orthologs not detected by classical full-length approaches, thus showing the interest of combining both types of strategies.

Another hassle of focusing on gene-level orthology is that, in Eukaryotes, a single gene may be transcribed into several isoforms with different exons combinations. This process, called alternative splicing, is especially prominent in vertebrates (Keren et al. 2010). Its functional implication is debated, but it has been shown for particular genes that different isoforms may have different tissue expression and even sometimes produce proteins with antagonist cellular functions (Wang et al. 2008). This has direct implications on the way orthology is used to transfer function between genes, as two orthologous genes could display different splicing patterns and even two orthologous genes with orthologous exons may have substantially different transcripts. Integrating homology between alternative transcripts of orthologs will provide additional information on whether an evolutionary conserved isoform is more likely to be functional, and whether observations made in a model species on a particular isoform are likely to be applicable to other species.

Assessing orthology between alternative transcripts often relies on two conditions (Blanquart et al. 2016). Indeed, transcripts are orthologous if (1) they are transcripts of orthologous genes and (2) their exons are similar enough to assume they are orthologous and appear in the same order in the gene sequence. The first condition is a classical orthology inference problem. The second condition may be determined by spliced sequence alignment, using an exon-aware alignment method (Kapustin et al. 2008; Gotoh 2008; Sharma et al. 2016; Jammali et al. 2019). Transcript orthology prediction has been successfully employed to identify orthologous isoforms between the gene repertoires of mouse and human (Zambelli et al. 2010). Applying it to more species is trickier since it cannot be done with pairwise relations and requires the construction of gene trees, which is computationally demanding. Nonetheless, it has been used to study multiple gene families, mapping events of isoform gains and losses to the branches of the trees (Christinat and Moret 2012; Jammali et al. 2019). Nevertheless, one must still be cautious when using isoform orthology determination and ensure that expression of both isoforms can be detected through experimental means in the species of interest, to avoid the pitfalls of erroneous annotation transfer.



As can be seen, despite the major advances made in recent years in orthology inference and resources, there is still a long way to go in the quest for orthologs. The practical and conceptual challenges are numerous and will require the efforts of the entire comparative genomics community to invent new solutions. Substantial progress will be needed both in the development of new indicators of proteome quality and for the formal representation of orthology relationships at different granularity levels.

**Acknowledgements** The authors thank Julie Thompson for critical reading of the manuscript. The authors are also grateful to the anonymous referees for their useful suggestions.

## References

- Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249. <https://doi.org/10.1038/nmeth0410-248>
- Alliance of Genome Resources Consortium (2020) Alliance of genome resources portal: unified model organism research platform. *Nucleic Acids Res* 48:D650–D658. <https://doi.org/10.1093/nar/gkz813>
- Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* 8:e1002514. <https://doi.org/10.1371/journal.pcbi.1002514>
- Altenhoff AM, Gil M, Gonnet GH, Dessimoz C (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE* 8:e53786. <https://doi.org/10.1371/journal.pone.0053786>
- Altenhoff AM, Boeckmann B, Capella-Gutierrez S et al (2016) Standardized benchmarking in the quest for orthologs. *Nat Methods* 13:425–430. <https://doi.org/10.1038/nmeth.3830>
- Altenhoff AM, Glover NM, Train C-M et al (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* 46:D477–D485. <https://doi.org/10.1093/nar/gkx1019>
- Altenhoff AM, Glover NM, Dessimoz C (2019) Inferring orthology and paralogy. In: Anisimova M (ed) *Evolutionary genomics: statistical and computational methods*. Springer, New York, NY, pp 149–175
- Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402. <https://doi.org/10.1093/nar/25.17.3389>
- Amar D, Frades I, Danek A et al (2014) Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. *BMC Plant Biol* 14:329. <https://doi.org/10.1186/s12870-014-0329-9>
- Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 25:25–29. <https://doi.org/10.1038/75556>
- Blanquart S, Varré J-S, Guertin P et al (2016) Assisted transcriptome reconstruction and splicing orthology. *BMC Genomics* 17:786. <https://doi.org/10.1186/s12864-016-3103-6>
- Brockhurst MA, Harrison E, Hall JPJ et al (2019) The ecology and evolution of pangenomes. *Curr Biol* CB 29:R1094–R1103. <https://doi.org/10.1016/j.cub.2019.08.012>
- Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60. <https://doi.org/10.1038/nmeth.3176>

- Buljan M, Bateman A (2009) The evolution of protein domain families. *Biochem Soc Trans* 37:751–755. <https://doi.org/10.1042/BST0370751>
- Burki F, Roger AJ, Brown MW, Simpson AGB (2020) The new tree of eukaryotes. *Trends Ecol Evol* 35:43–55. <https://doi.org/10.1016/j.tree.2019.08.008>
- Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. *BMC Bioinform* 10:421. <https://doi.org/10.1186/1471-2105-10-421>
- Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ (2004) The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res* 14:802–811. <https://doi.org/10.1101/gr.2195604>
- Chen X, Zhang J (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol* 8:e1002784. <https://doi.org/10.1371/journal.pcbi.1002784>
- Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34:D363–D368. <https://doi.org/10.1093/nar/gkj123>
- Christinat Y, Moret BME (2012) Inferring transcript phylogenies. *BMC Bioinform* 13(Suppl 9):S1. <https://doi.org/10.1186/1471-2105-13-s9-s1>
- Crawford NG, Faircloth BC, McCormack JE et al (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Lett* 8:783–786. <https://doi.org/10.1098/rsbl.2012.0331>
- Darby CA, Stolzer M, Ropp PJ et al (2017) Xenolog classification. *Bioinformatics* 33:640–649. <https://doi.org/10.1093/bioinformatics/btw686>
- Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* 9:157. <https://doi.org/10.1186/1471-2148-9-157>
- Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comput Biol* 7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- El-Gebali S, Mistry J, Bateman A et al (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427–D432. <https://doi.org/10.1093/nar/gky995>
- Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16:157. <https://doi.org/10.1186/s13059-015-0721-2>
- Emms DM, Kelly S (2018) STAG: species tree inference from all genes. *bioRxiv* 267914. <https://doi.org/10.1101/267914>
- Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238. <https://doi.org/10.1186/s13059-019-1832-y>
- Feng B, Zhou L, Tang J (2017) Ancestral genome reconstruction on whole genome level. *Curr Genomics* 18:306–315. <https://doi.org/10.2174/1389202918666170307120943>
- Fernández-Breis JT, Chiba H, Legaz-García MDC, Uchiyama I (2016) The orthology ontology: development and applications. *J Biomed Semant* 7:34. <https://doi.org/10.1186/s13326-016-0077-x>
- Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst Zool* 19:99–113
- Fitch WM (2000) Homology a personal view on some of the problems. *Trends Genet TIG* 16:227–231. [https://doi.org/10.1016/s0168-9525\(00\)02005-9](https://doi.org/10.1016/s0168-9525(00)02005-9)
- Force A, Lynch M, Pickett FB et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Forslund K, Pekkari I, Sonnhammer ELL (2011) Domain architecture conservation in orthologs. *BMC Bioinform* 12:326. <https://doi.org/10.1186/1471-2105-12-326>
- Gabaldón T, Dessimoz C, Huxley-Jones J et al (2009) Joining forces in the quest for orthologs. *Genome Biol* 10:403. <https://doi.org/10.1186/gb-2009-10-9-403>
- Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature* 451:704–707. <https://doi.org/10.1038/nature06510>
- Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief Bioinform* 12:449–462

- Gilks WR, Audit B, De Angelis D et al (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* 18:1641–1649. <https://doi.org/10.1093/bioinformatics/18.12.1641>
- Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: what are they and how do we infer them? *Trends Plant Sci* 21:609–621. <https://doi.org/10.1016/j.tplants.2016.02.005>
- Gotoh O (2008) Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics* 24:2438–2444. <https://doi.org/10.1093/bioinformatics/btn460>
- Gray GS, Fitch WM (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol Biol Evol* 1:57–66. <https://doi.org/10.1093/oxfordjournals.molbev.a040298>
- Grossetête S, Labedan B, Lespinet O (2010) FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics* 11:81. <https://doi.org/10.1186/1471-2164-11-81>
- Gumulya Y, Gillam EMJ (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the “retro” approach to protein engineering. *Biochem J* 474:1–19. <https://doi.org/10.1042/BCJ20160507>
- Hecker N, Sharma V, Hiller M (2019) Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proc Natl Acad Sci* 116:3036–3041. <https://doi.org/10.1073/pnas.1818504116>
- Henricson A, Forslund K, Sonnhammer ELL (2010) Orthology confers intron position conservation. *BMC Genomics* 11:412. <https://doi.org/10.1186/1471-2164-11-412>
- Herrero J, Muffato M, Beal K et al (2016) Ensembl comparative genomics resources. *Database J Biol Databases Curation*. <https://doi.org/10.1093/database/baw053>
- Hu X, Friedberg I (2019) SwiftOrtho: a fast, memory-efficient, multiple genome orthology classifier. *GigaScience* 8. <https://doi.org/10.1093/gigascience/giz118>
- Hu Y, Flockhart I, Vinayagam A et al (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinform* 12:357. <https://doi.org/10.1186/1471-2105-12-357>
- Huang X, Albou L-P, Mushayahama T et al (2019) Ancestral genomes: a resource for reconstructed ancestral genes and genomes across the tree of life. *Nucleic Acids Res* 47:D271–D279. <https://doi.org/10.1093/nar/gky1009>
- Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T (2007) The human phylome. *Genome Biol* 8:R109. <https://doi.org/10.1186/gb-2007-8-6-r109>
- Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP et al (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* 42:D897–D902. <https://doi.org/10.1093/nar/gkt1177>
- Huerta-Cepas J, Szklarczyk D, Forslund K et al (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44:D286–D293. <https://doi.org/10.1093/nar/gkv1248>
- Huerta-Cepas J, Forslund K, Coelho LP et al (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 34:2115–2122. <https://doi.org/10.1093/molbev/msx148>
- Hug LA, Baker BJ, Anantharaman K et al (2016) A new view of the tree of life. *Nat Microbiol* 1:16048. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett* 426:1–5. [https://doi.org/10.1016/s0014-5793\(98\)00276-2](https://doi.org/10.1016/s0014-5793(98)00276-2)
- Jammali S, Aguilar J-D, Kuitche E, Ouangraoua A (2019) SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups. *BMC Bioinform* 20:133. <https://doi.org/10.1186/s12859-019-2647-2>
- Jensen LJ, Julien P, Kuhn M et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 36:D250–D254. <https://doi.org/10.1093/nar/gkm796>

- Jim K, Parmar K, Singh M, Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Res* 14:109–115. <https://doi.org/10.1101/gr.1586704>
- Kapustin Y, Souvorov A, Tatusova T, Lipman D (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct* 3:20. <https://doi.org/10.1186/1745-6150-3-20>
- Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J Roy Soc Interface* 5:151–170. <https://doi.org/10.1098/rsif.2007.1047>
- Kent WJ, Baertsch R, Hinrichs A et al (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100:11484–11489. <https://doi.org/10.1073/pnas.1932072100>
- Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11:345–355. <https://doi.org/10.1038/nrg2776>
- Khan FI, Wei D-Q, Gu K-R et al (2016) Current updates on computer aided protein modeling and designing. *Int J Biol Macromol* 85:48–62. <https://doi.org/10.1016/j.ijbiomac.2015.12.072>
- Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101:573–576. [https://doi.org/10.1016/S0092-8674\(00\)80867-3](https://doi.org/10.1016/S0092-8674(00)80867-3)
- Koonin EV, Fedorova ND, Jackson JD et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 5:R7
- Kriventseva EV, Kuznetsov D, Tegenfeldt F et al (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 47:D807–D811. <https://doi.org/10.1093/nar/gky1053>
- Kryuchkova-Mostacci N, Robinson-Rechavi M (2015) Tissue-specific evolution of protein coding genes in human and mouse. *PLoS ONE* 10:e0131673. <https://doi.org/10.1371/journal.pone.0131673>
- Laetsch DR, Blaxter ML (2017) KinFin: software for taxon-aware analysis of clustered protein sequences. *G3 Bethesda Md* 7:3349–3357. <https://doi.org/10.1534/g3.117.300233>
- Land M, Hauser L, Jun S-R et al (2015) Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 15:141–161. <https://doi.org/10.1007/s10142-015-0433-4>
- Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. *Trends Genet TIG* 25:107–110. <https://doi.org/10.1016/j.tig.2008.12.004>
- Lecompte O, Thompson JD, Plewniak F et al (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* 270:17–30. [https://doi.org/10.1016/s0378-1119\(01\)00461-9](https://doi.org/10.1016/s0378-1119(01)00461-9)
- Leelananda SP, Lindert S (2016) Computational methods in drug discovery. *Beilstein J Org Chem* 12:2694–2718. <https://doi.org/10.3762/bjoc.12.267>
- Li L, Stoekert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189. <https://doi.org/10.1101/gr.1224503>
- Linard B, Thompson JD, Poch O, Lecompte O (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinform* 12:11. <https://doi.org/10.1186/1471-2105-12-11>
- Liu T, Tang GW, Capriotti E (2011) Comparative modeling: the state of the art and protein drug target structure prediction. *Comb Chem High Throughput Screen* 14:532–547. <https://doi.org/10.2174/138620711795767811>
- Mi H, Dong Q, Muruganujan A et al (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. *Nucleic Acids Res* 38:D204–D210. <https://doi.org/10.1093/nar/gkp1019>
- Mi H, Muruganujan A, Ebert D et al (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 47:D419–D426. <https://doi.org/10.1093/nar/gky1038>
- Mukherjee S, Stamatis D, Bertsch J et al (2019) Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Res* 47:D649–D659. <https://doi.org/10.1093/nar/gky977>

- NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44:D7–D19. <https://doi.org/10.1093/nar/gkv1290>
- Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol* 7:e1002073. <https://doi.org/10.1371/journal.pcbi.1002073>
- Nevers Y, Prasad MK, Poidevin L et al (2017) Insights into ciliary genes and evolution from multi-level phylogenetic profiling. *Mol Biol Evol* 34:2016–2034. <https://doi.org/10.1093/molbev/msx146>
- Nevers Y, Kress A, Defosset A et al (2019) OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Res* 47:D411–D418. <https://doi.org/10.1093/nar/gky1068>
- Overbeek R, Fonstein M, D'Souza M et al (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 96:2896–2901. <https://doi.org/10.1073/pnas.96.6.2896>
- Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067. <https://doi.org/10.1093/bioinformatics/btm071>
- Parra G, Bradnam K, Ning Z et al (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res* 37:289–297. <https://doi.org/10.1093/nar/gkn916>
- Pellegrini M, Marcotte EM, Thompson MJ et al (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96:4285–4288. <https://doi.org/10.1073/pnas.96.8.4285>
- Penel S, Arigon A-M, Dufayard J-F et al (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinform* 10(Suppl 6):S3. <https://doi.org/10.1186/1471-2105-10-S6-S3>
- Pereira C, Denise A, Lespinet O (2014) A meta-approach for improving the prediction and the functional annotation of ortholog groups. *BMC Genomics* 15(Suppl 6):S16. <https://doi.org/10.1186/1471-2164-15-S6-S16>
- Persson E, Kaduk M, Forslund SK, Sonnhammer ELL (2019) Domainoid: domain-oriented orthology inference. *BMC Bioinform* 20:523. <https://doi.org/10.1186/s12859-019-3137-2>
- Peterson ME, Chen F, Saven JG et al (2009) Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci Publ Protein Soc* 18:1306–1315. <https://doi.org/10.1002/pro.143>
- Petersen M, Meusemann K, Donath A et al (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinform* 18:111. <https://doi.org/10.1186/s12859-017-1529-8>
- Pryszcz LP, Huerta-Cepas J, Gabaldón T (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res* 39:e32. <https://doi.org/10.1093/nar/gkq953>
- Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314:1041–1052. <https://doi.org/10.1006/jmbi.2000.5197>
- Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinform* 9:518. <https://doi.org/10.1186/1471-2105-9-518>
- Rouard M, Guignon V, Aluome C et al (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res* 39:D1095–D1102. <https://doi.org/10.1093/nar/gkq811>
- Santpere G, Lopez-Valenzuela M, Petit-Marty N et al (2016) Differences in molecular evolutionary rates among microRNAs in the human and chimpanzee genomes. *BMC Genomics* 17:528. <https://doi.org/10.1186/s12864-016-2863-3>
- Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform* 12:485–488. <https://doi.org/10.1093/bib/bbr025>
- Schoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 5:e1000605. <https://doi.org/10.1371/journal.pcbi.1000605>

- Schreiber F, Sonnhammer ELL (2013) Hieranoid: hierarchical orthology inference. *J Mol Biol* 425:2072–2081. <https://doi.org/10.1016/j.jmb.2013.02.018>
- Schreiber F, Patricio M, Muffato M et al (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res* 42:D922–D925. <https://doi.org/10.1093/nar/gkt1055>
- Schwartz S, Kent WJ, Smit A et al (2003) Human-mouse alignments with BLASTZ. *Genome Res* 13:103–107. <https://doi.org/10.1101/gr.809403>
- Sharma V, Elghafari A, Hiller M (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Res* 44:e103. <https://doi.org/10.1093/nar/gkw210>
- Sjolander K, Datta RS, Shen Y, Shoffner GM (2011) Ortholog identification in the presence of domain architecture rearrangement. *Brief Bioinform* 12:413–422. <https://doi.org/10.1093/bib/bbr036>
- Škunca N, Dessimoz C (2015) Phylogenetic profiling: how much input data is enough? *PLoS ONE* 10:e0114701. <https://doi.org/10.1371/journal.pone.0114701>
- Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197. [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)
- Sonnhammer ELL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet TIG* 18:619–620. [https://doi.org/10.1016/s0168-9525\(02\)02793-2](https://doi.org/10.1016/s0168-9525(02)02793-2)
- Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43:D234–D239. <https://doi.org/10.1093/nar/gku1203>
- Sonnhammer ELL, Gabaldón T, Sousa da Silva AW et al (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics* 30:2993–2998. <https://doi.org/10.1093/bioinformatics/btu492>
- Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet* 16:472–482. <https://doi.org/10.1038/nrg3962>
- Spang A, Saw JH, Jørgensen SL et al (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521:173–179. <https://doi.org/10.1038/nature14447>
- Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026–1028. <https://doi.org/10.1038/nbt.3988>
- Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18:92–99. <https://doi.org/10.1093/bioinformatics/18.1.92>
- Straub K, Merkl R (2019) Ancestral sequence reconstruction as a tool for the elucidation of a stepwise evolutionary adaptation. *Methods Mol Biol Clifton NJ* 1851:171–182. [https://doi.org/10.1007/978-1-4939-8736-8\\_9](https://doi.org/10.1007/978-1-4939-8736-8_9)
- Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet TIG* 25:210–216. <https://doi.org/10.1016/j.tig.2009.03.004>
- Sutphin GL, Mahoney JM, Sheppard K et al (2016) WORMHOLE: novel least diverged ortholog prediction through machine learning. *PLoS Comput Biol* 12:e1005182. <https://doi.org/10.1371/journal.pcbi.1005182>
- Szklarczyk D, Gable AL, Lyon D et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47:D607–D613. <https://doi.org/10.1093/nar/gky1131>
- Tabach Y, Golan T, Hernández-Hernández A et al (2013) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol Syst Biol* 9:692. <https://doi.org/10.1038/msb.2013.50>
- Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637. <https://doi.org/10.1126/science.278.5338.631>
- The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res* 47:D330–D338. <https://doi.org/10.1093/nar/gky1055>
- The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47:D506–D515. <https://doi.org/10.1093/nar/gky1049>

- Uchiyama I, Mihara M, Nishide H et al (2019) MGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res* 47:D382–D389. <https://doi.org/10.1093/nar/gky1054>
- Van Bel M, Diels T, Vancaester E et al (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res* 46:D1190–D1196. <https://doi.org/10.1093/nar/gkx1002>
- Van de Peer Y, Mizrahi E, Marchal K (2017) The evolutionary significance of polyploidy. *Nat Rev Genet* 18:411–424. <https://doi.org/10.1038/nrg.2017.26>
- Vaser R, Adusumalli S, Leng SN et al (2016) SIFT missense predictions for genomes. *Nat Protoc* 11:1–9. <https://doi.org/10.1038/nprot.2015.123>
- Vilella AJ, Severin J, Ureta-Vidal A et al (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19:327–335. <https://doi.org/10.1101/gr.073585.107>
- Walhout AJ, Boulton SJ, Vidal M (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* Chichester Engl 17:88–94. [https://doi.org/10.1002/1097-0061\(20000630\)17:2%3c88::AID-YEA20%3e3.0.CO;2-Y](https://doi.org/10.1002/1097-0061(20000630)17:2%3c88::AID-YEA20%3e3.0.CO;2-Y)
- Wang ET, Sandberg R, Luo S et al (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476. <https://doi.org/10.1038/nature07509>
- Waterhouse RM, Seppey M, Simão FA et al (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 35:543–548. <https://doi.org/10.1093/molbev/msx319>
- Whiteside MD, Winsor GL, Laird MR, Brinkman FSL (2013) OrthoLugeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. *Nucleic Acids Res* 41:D366–D376. <https://doi.org/10.1093/nar/gks1241>
- Wolfe K (2000) Robustness—it's not where you think it is. *Nat Genet* 25:3–4. <https://doi.org/10.1038/75560>
- Wu Y-C, Rasmussen MD, Kellis M (2012) Evolution at the subgene level: domain rearrangements in the drosophila phylogeny. *Mol Biol Evol* 29:689–705. <https://doi.org/10.1093/molbev/msr222>
- Zambelli F, Pavesi G, Gissi C et al (2010) Assessment of orthologous splicing isoforms in human and mouse orthologous genes. *BMC Genomics* 11:534. <https://doi.org/10.1186/1471-2164-11-534>
- Zaremba-Niedzwiedzka K, Caceres EF, Saw JH et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541:353–358. <https://doi.org/10.1038/nature21031>
- Zmasek CM, Eddy SR (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinform* 3:14. <https://doi.org/10.1186/1471-2105-3-14>

# Chapter 10

## Prehistoric Stone Projectile Points and Technological Convergence



Michael J. O'Brien  and George R. McGhee

**Abstract** Stone tools are found throughout the archaeological record left by humans and their ancestors beginning as much as 2.6–3.4 million years ago. Given the nearly ubiquitous use of stone tools by hominins, their study is an important line of inquiry for shedding light on questions of evolution and behavior. Because they were parts of past phenotypes, stone tools were shaped by the same evolutionary processes as were the somatic (bodily) features of their users. One evolutionary result of these processes is convergence—the appearance of similar forms in independent lineages that result from functional or developmental constraints. With respect to stone tools, identifying cases of convergence is particularly important because similarities in form are often used to suggest historical connections among prehistoric groups. Identifying cases of convergence would refute hypotheses that otherwise would suggest some degree of physical or cultural connection.

### 10.1 Introduction

Even before Darwin (1859) wrote *On the Origin of Species*, naturalists made a distinction between biological *divergence*—the splitting of two or more lines related through a common ancestor—and biological *convergence*—two unrelated lines landing on the same point on an adaptive landscape. Similarly, naturalists distinguished between what later would be called *analogous* traits and *homologous* traits. Analogous traits are those that two or more organisms possess that, although they might serve similar purposes, did not evolve through common ancestry. Rather, they result from convergence. For example, birds and bats both have wings, and those traits share properties in common, yet we classify birds and bats in two widely separate taxonomic groups because birds and bats are only distantly related. This is because these two large

---

M. J. O'Brien (✉)

Office of the Provost, Texas A&M University—San Antonio, San Antonio, TX 78253, USA  
e-mail: [mike.obrien@tamusa.edu](mailto:mike.obrien@tamusa.edu)

G. R. McGhee

Department of Earth and Planetary Sciences, Rutgers University, Piscataway, NJ 08854, USA  
e-mail: [mcghee@eps.rutgers.edu](mailto:mcghee@eps.rutgers.edu)



groups diverged from a common vertebrate ancestor long before either one of them developed wings. Therefore, wings are of no utility in understanding patterns of descent because they evolved independently in the two lineages after they diverged. Conversely, homologous traits *are* useful for tracking continuity resulting from inheritance because they are holdovers from the time when two lineages were historically a single lineage and then diverged. For example, all mammals have a vertebral column, as do animals placed in other categories. The presence of vertebrae is one criterion that we use to place organisms in the subphylum Vertebrata. The vertebral column is a trait shared by mammals, birds, reptiles, and some fishes, and it suggests that at some remote time in the past, organisms in these groups shared a common ancestor.

Anthropologists and archaeologists have long understood the distinction between cultural divergence and convergence, although they have not always been able to consistently differentiate between the two. The means for doing began to be applied in biology and paleobiology with the introduction of cladistics into the English-speaking world in the 1960s (Hennig 1966), but it would be two decades before the method made its way into cultural studies (e.g., Foley 1987) and then only sporadically. It wasn't until the mid-1990s that cladistics began to be applied more broadly across cultural domains, especially in linguistics and cultural anthropology (e.g., Holden and Mace 1997; Mace and Pagel 1994), and shortly thereafter in archaeology (e.g., Collard and Shennan 2000; O'Brien and Lyman 2003; O'Brien et al. 2001).

The use of cladistics in archaeology came about as researchers began to view the archaeological record in more or less the same way that paleobiologists view a fossil bed: as a population of objects that represent the hard parts of past phenotypes. Because they were phenotypic, the objects were subject to the same evolutionary processes—selection, drift, and recombination—as were the somatic features of their makers and users (Leonard and Jones 1987). That things such as stone tools are phenotypic is nonproblematic to most biologists (e.g., Bonner 1988; Dawkins 1982; Odling-Smee and Turner 2011; Turner 2000, 2012), who routinely view a bird's nest, a beaver's dam, or a chimpanzee's twig tools as phenotypic traits. That view certainly is not problematic to paleobiologists, who have to rely on the hard parts of phenotypes (shells, for example) to study the evolution of extinct organisms and their lineages.

Once we accept that point, we can begin to talk about selection and drift and how they shaped the variation that shows up in the archaeological record—variation that resulted from convergence as well as divergence (O'Brien 2019). Our goal in this chapter is to outline how archaeologists have approached the problem of identifying convergence and to propose a more-integrated approach that makes maximum use of various analytical tools that currently exist. Throughout our discussion we focus primarily on stone tools, where identifying cases of convergence is particularly important because similarities in form have been, and, unfortunately, in many cases continue to be, used uncritically to make historical connections among prehistoric groups, many of which have later been shown to be nonexistent.

### 10.1.1 A Brief Background to the Problem

To archaeologists working throughout much of the twentieth century, culture was viewed as something that evolved, but any similarity between biological and cultural evolution was seen as strictly metaphorical (Kroeber 1923), with the former being linked to genetic transmission and the latter to something entirely different (Brew 1946). The standard view was that any attempt to link one of Darwin's mechanisms for change—natural selection, for example—to the evolution of culture was little more than misapplied biology. Steward (1941, p. 367) pointed this out quite forcefully: “It is apparent... that strict adherence to a method drawn from biology inevitably fails to take into account the distinctively cultural and unbiological fact of blends and crosses between essentially unlike types.... A taxonomic scheme cannot indicate this fact without becoming mainly a list of exceptions.” We will return to the issue of “blends” and “crosses” in Sect. 3.3.

Despite their distaste for biological evolution as a framework for understanding cultural evolution, archaeologists showed a knowledge of and appreciation for the distinction between convergence and divergence, although they were unable to make the distinction in analytical terms. Kroeber (1931, pp. 152–153) had this to say on the subject:

There are cases in which it is not a simple matter to decide whether the totality of traits points to a true relationship or to secondary convergence.... Yet few biologists would doubt that sufficiently intensive analysis of structure will ultimately solve such problems of descent.... There seems no reason why on the whole the same cautious optimism should not prevail in the field of culture; why homologies should not be positively distinguishable from analogies when analysis of the whole of the phenomena in question has become truly intensive. That such analysis has often been lacking but judgments have nevertheless been rendered, does not invalidate the positive reliability of the method.

Although he was clear that there are two forms of similarity, Kroeber was unclear as to how one might distinguish between them. He pointed out that identifying “similarities [that] are specific and structural and not merely superficial... has long been the accepted method in evolutionary and systematic biology” (Kroeber 1931, p. 151), but he offered no opinion as to how to separate them beyond undertaking a “sufficiently intensive analysis of structure.” Kroeber was correct: An intensive analysis of structure, especially a detailed *comparative* analysis, is critical to being able to make the distinction, but again, he did not offer any advice on how to do that. As a result, Kroeber—and he was by no means alone—landed on the default option: Formal similarities between sets of artifacts must signal *some* kind of relationship, either an ancestor–descendant relationship or one derived through ethnologically documented mechanisms such as diffusion and enculturation (Lyman et al. 1997).

Gordon Willey (1953, p. 363) did not waffle on the matter, declaring axiomatically that “typological similarity *is* an indicator of cultural relatedness (and this is surely axiomatic to archeology), [and thus] such relatedness carries with it implications of a common or similar history” (emphasis added). This axiom, however, falls prey to a

caution raised by paleontologist George Gaylord Simpson (1961), using monozygotic twins as an example: They are twins not because they are similar; rather, they are similar because they are twins and thus share a common history. There is a big difference between the two (O'Brien and Lyman 2000).

The default option—formal similarity signals relationship—continued to dominate archaeology, and the number of articles and monographs emphasizing diffusion and migration as explanatory devices continued to increase throughout the twentieth century. As Rowe (1966, p. 334) noted, however, most accounts were nothing more than poorly concocted just-so stories: “We are now being subjected in archaeological meetings to ever more strident claims that Mesoamerican culture was derived from China or southeast Asia, early Ecuadorian culture from Japan, Woodland culture from Siberia, Peruvian culture from Mesoamerica, and so forth. In the science-fiction world of the diffusionists, a dozen similarities of detail prove cultural contact, and time, distance, and the difficulties of navigation are assumed to be irrelevant.”

We do not have to go back into the twentieth century to find archaeological examples of the conflation of divergence and convergence. Less than a decade ago, Dennis Stanford and Bruce Bradley published *Across Atlantic Ice: The Origin of America's Clovis Culture* (Stanford and Bradley 2012), which was the latest iteration of their proposal that North America was first colonized by people from Europe rather than from East Asia, as is commonly accepted in North American archaeology (Moreno-Mayar et al. 2018). Stanford and Bradley argued that Solutrean people from the Iberian Peninsula and southern France used some sort of boat or raft to make their way across the North Atlantic and into North America during the Last Glacial Maximum, some 20,000–24,000 years ago. Under this “Solutrean hypothesis,” the 6000-km journey was made possible by a continuous ice shelf that provided fresh water and a stable food supply. In its initial formulation, the hypothesis was based primarily on similarities between the stone tools and production techniques of Solutrean people from Western Europe, which date about 23,500–18,000 calibrated radiocarbon years before present (cal BP) (Straus 2005), and the tools and techniques of North American Clovis people, which date about 13,300–12,800 cal BP in western North America and ca. 12,800–12,200 cal BP in eastern North America (Gingerich 2011; Haynes 2015; Miller et al. 2014).

Flaws in the Solutrean hypothesis were immediately obvious. For one thing, the multiple-thousand-year gap between Solutrean and Clovis made an ancestor–descendant relationship highly improbable, meaning that similarities in tool design were instead the result of convergence: unrelated populations of prehistoric flintknappers found similar solutions to similar adaptive problems (Straus 2000). To deal with the large chronological gap, Stanford and Bradley shifted their focus from similarities between Solutrean and Clovis to supposed similarities among Solutrean, Clovis, and pre-Clovis tool types and production techniques (Bradley and Stanford 2004; Stanford and Bradley 2002). This was an unfortunate modification to their proposal because the pre-Clovis

dates used by Stanford and Bradley—all of which were from highly questionable contexts—actually predate the Solutrean (O’Brien et al. 2014a, b). This would suggest that the traits appeared first in North America and then were carried to Europe. This, of course, is implausible. We will come back periodically to the Solutrean example because it is an excellent case of how what we propose here can help us escape the conundrum of similarity automatically signaling relatedness.

## 10.2 A Way Forward

As a prelude to our discussion, let us look at what archaeologist David Clarke (1968) had to say on the important distinction between two terms, *phyletic* and *phenetic*. Understanding both the distinction and the concepts behind the terms offers us a way forward in being able to distinguish between divergence and convergence:

One of the fundamental problems that the archaeologist repeatedly encounters is the assessment of whether a set of archaeological entities are connected by a direct cultural relationship linking their generators or whether any affinity between the set is based on more general grounds. This problem usually takes the form of an estimation of the degree of affinity or similarity between the entities and then an argument as to whether these may represent a genetic and phyletic lineage or merely a phenetic and non-descent connected affinity. (p. 211)

Note that the terms “phyletic”—“phylogenetic” is a more appropriate term—and “phenetic” are both grounded in the concept of similarity, but the former signifies a descent-related affinity—one person, population, or object being related to another one (or more)—whereas the latter has nothing to do with descent. Despite recognizing the key distinction, Clarke was unable to propose how to separate instances of convergence from instances of divergence except to rely on degree of similarity: The more similar, the more related two things are. Again, this missed Simpson’s (1961) caution. We see four interrelated means of heeding Simpson’s caution and determining whether prehistoric stone tools are the result of convergence or divergence: (1) experimental replication, (2) consideration of functional and developmental constraints, (3) phylogenetic—as opposed to phenetic—analysis, and (4) use of more-precise language with respect to identifying kinds of convergence. We examine each of these below.

### 10.2.1 *Experimental Replication*

Replication—here of flaked-stone tools—is the act of creating specimens for one or more experimental purposes: (1) to create a framework for generating hypotheses about tool manufacture; (2) to test a specific hypothesis about certain parameters of stone-tool technology; and (3) to validate quantitative methods that will be used to

study archaeological tools and their by-products of manufacture (Eren et al. 2016). Hypotheses and their predictions determine the variables required for an experiment, including such things as the sample size of participants or specimens, measurement, and test protocols, whether the experiment is a blind test, and the quantitative methods and statistical analyses that are selected (Eren et al. 2016).

Archaeologists who use replication experiments face several challenges (Kelly 1994), including mistakenly using their intuitive knowledge of stone-tool replication to extend to a belief that their knowledge somehow gives them unique insights into such things as prehistoric foraging behavior and adaptation (Thomas 1986). Perhaps in reaction to this latter behavior, some archeologists dismiss the usefulness of any experiment based on stone-tool replication. Both viewpoints stem from a poor articulation of the principle of uniformitarianism (Eren et al. 2016). Under the first viewpoint, the principle of uniformitarianism is exaggerated to such an extent that a scientific framework no longer becomes necessary to test hypotheses because the knapper simply “knows” the past because he or she is “reproducing” it. The second viewpoint ignores the fact that stone breaks the same way today as it did in the past and possesses the same physical properties as it did in the past—sharp cutting edges, durability, shape, and so on—which provides a uniformitarian link that is exploitable scientifically (Eren et al. 2016) (Fig. 1). Now, instead of adopting hyperdiffusionist explanations of the archaeological record, where prehistoric people are equated with



**Fig. 1** Stone-tool production involves reducing a large piece of flint or other stone material through percussion flaking—shown here—and/or pressure flaking. Flakes of various sizes are removed in tool production, some of which themselves can be further modified for use as tools. Courtesy Metin Eren

their technology, we can begin to better understand that similar technologies can be developed independently—that, for example, tool types on opposite sides of an ocean and separated by thousands of years share similar production processes.

Returning to the Solutrean hypothesis, Stanford and Bradley argued that overshot flaking, a reduction technique in which long flakes are struck from prepared edges of a biface and travel across the face and remove a portion of the opposite margin, was used to produce both Clovis and Solutrean points. Overshot flaking is a difficult technique that few modern knappers have mastered (Eren et al. 2013, 2014), but Stanford and Bradley were convinced that the technique was intentionally used by Solutrean and Clovis peoples because of its presumed advantages, especially for rapidly thinning stone bifaces. Stanford and Bradley (2012, pp. 28, 157) made two other claims: first, that there was “clear archaeological evidence of widespread use” of overshot flaking by Solutrean and Clovis knappers, and second, that the “level of correspondence between the two technologies is amazing,” such that “even the details of flaking are virtually identical.” They then argued that because the intentional use of a complex, difficult strategy is unlikely to occur by chance, its presence in two separate groups suggests that it was unlikely to have been independently invented.

Metin Eren—a master flint knapper—and colleagues, one of whom was also a master flint knapper, used replication experiments and quantitative analysis of the archeological record to evaluate Stanford and Bradley’s claims (Eren et al. 2013, 2014). They found that, unlike in Stanford and Bradley’s proposal, overshot flaking is most easily explained as a technological by-product of reducing stone rather than a complex knapping strategy. They found that overshot flaking is no more efficient at thinning a biface than non-overshot flaking and that there is no frequent occurrence of overshot evidence at Clovis sites and no published data on the frequency or regularity of overshot flaking at Solutrean sites. Further, they pointed out that Stanford and Bradley reported but a single flake for overshot flaking in their 14 purported pre-Clovis assemblages. It is difficult to demonstrate historical relatedness when using only a single flake.

Embedded in experimental replication is the concept of fidelity—how faithfully something is replicated, or copied—although a distinction should be made between two kinds of copying: *imitation*, in which the form of an action is copied, and *emulation*, in which the result of an action sequence is copied. With respect to the manufacture of a Clovis point, for example, there is a clear distinction between imitation—understanding the actions necessary to produce a point—and emulation—trying to produce a point without understanding the necessary actions (and their correct sequence). Stone-reduction sequences are complex procedures that require a significant amount of investment in terms of time and energy to learn effectively (Geribàs et al. 2010; Stout 2011), and Clovis-point production is no exception (Bradley et al. 2010). Fluting (Fig. 2), for example, is a challenging technology to master, occurring after a point is already thinned to approximately 7.5 mm (Thomas et al. 2017). That does not give the knapper much margin of error. This leads directly into a discussion of two kinds of constraints that knappers face.



**Fig. 2** Fluted points from North America. They were bifacially flaked (flaked on both sides) from any one of several cryptocrystalline stone types, such as chert, quartzite, and obsidian. The points are lanceolate in form, have parallel to slightly convex sides and concave bases, and exhibit a series of flake-removal scars—“flutes”—on one or both faces that extend from the base to about a third of the way to the tip. Experimental evidence suggests that the thinner base that results from fluting acts as a “shock absorber” that increases point robustness and the ability to withstand physical stress through stress redistribution and damage relocation (Story et al. 2019; Thomas et al. 2017). The specimen on the left is a Clovis point; the one in the middle is a Cumberland point; and the one on the right is a Folsom point. Photos courtesy Pete Bostrum and the Lithic Casting Lab

### 10.2.2 *Functional and Developmental Constraints*

Experimental replication allows for a better understanding of two important considerations with respect to stone-tool production, functional and developmental constraints (McGhee 2018a; see also McGhee 2011). With respect to the former, which is an *extrinsic* evolutionary constraint (McGhee 2007), given the same function, natural selection should produce the same tool form or production process to serve that function (McGhee 2011). With respect to chipped-stone technology, function may include tasks such as cutting, shooting, scraping, engraving, striking a large flake, or creating a series of parallel flake scars; attributes such as efficiency, effectiveness, portability, or durability; and social objectives such as costly signaling

or establishing group coherence (Eren et al. 2018). With respect to developmental constraint, which is an *intrinsic* evolutionary constraint (McGhee 2007), the different types of tool forms or production processes that humans can develop are limited (McGhee 2018b).

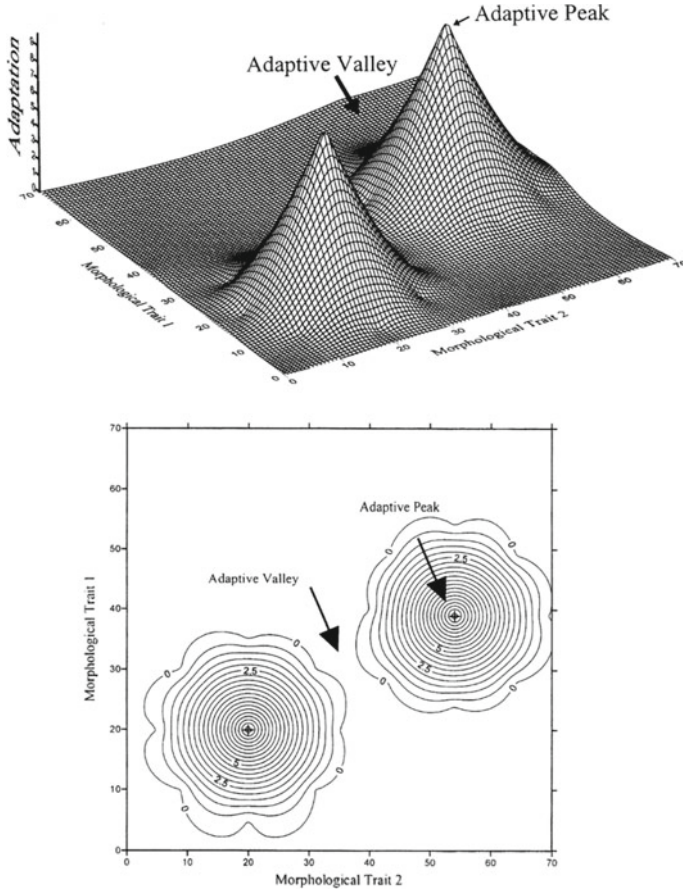
One way of modeling constraint is through the use of a heuristic that we will refer to as *theoretical morphospace*—a device that originated in evolutionary biology (McGhee 1999) but which is being increasingly used in archaeology (e.g., Charbonneau 2018; O'Brien and Bentley 2011; O'Brien et al. 2016). We can think of morphospace as a three-dimensional landscape that contains peaks of varying height, with height being a proxy for fitness, or adaptedness (McGhee 2007; O'Brien et al. 2016) (Fig. 3). Geneticist Sewall Wright (1932, 1988) introduced the metaphor of a fitness (adaptive) landscape to describe the possible mutational trajectories that lineages take (evolve) from genotypes that lie in regions of low fitness to regions of higher fitness (Kvitek and Sherlock 2011). We can borrow this metaphorical landscape and adapt its features so that the highest peak on the landscape corresponds to the optimal form of something, say, of an arrowhead, and lower peaks correspond to forms that, although not optimal, are good enough for the intended function at particular points in time (note the lower peak in the top half of Fig. 3). The landscape also contains adaptive valleys, which correspond to forms that yield negative fitness. An example of the latter would be a stone spear tip that is so thin that it consistently snaps on the slightest impact.

Fitness, then, is measured in terms of the success that one form exhibits relative to another, with success measured in terms of how often something is replicated (O'Brien et al. 2016). This can be determined through careful analysis of the archaeological record, provided that we have an objective means of classifying forms into one group, or class, as opposed to another (see below). Although the fitness of things such as stone tools as measured through differential replication is not *necessarily* linked to the fitness of humans—the propensity of individuals to live longer, have more offspring, and the like—there is considerable archaeological evidence that the relative fitness of one tool form over another *can* affect the relative fitness of the agents using the tools (Leonard 2001; Leonard and Jones 1987; Lyman and O'Brien 1998).

Any given form—a specific chipped-stone tool, for example—can be described by a set of measurements taken from that form—its length, its width, and so on. Each type of measurement can be considered as a dimension of form, and the total set of possible dimensions can be used to construct a hyperdimensional morphospace of possible form coordinates (McGhee 2011, 2018b, 2019). Each point within that theoretical morphospace represents a specific combination of form measurements that will produce the form coordinate for a hypothetical form. Convergence occurs when forms originally present in different regions of the morphospace evolve in such a way that they move to the same spatial region.

One simple theoretical morphospace is shown in Fig. 4, which represents a three-trait classification. One trait, height, has three possible states (1–3), depth also has three (A–C), and width has two (I and II). The regions of morphospace formed by the

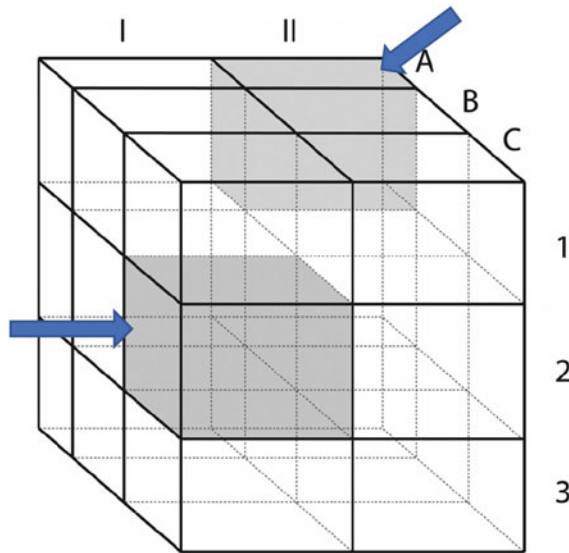




**Fig. 3** A hypothetical adaptive landscape, portrayed as both a three-dimensional grid (top) and a two-dimensional contour map (bottom). Topographic highs represent adaptive morphologies that function well in natural environments—and therefore are selected for—and topographic lows represent nonadaptive morphologies that function poorly—and therefore are selected against. On the contour map, the top of an adaptive peak is indicated by a plus sign, following the convention of Sewell Wright (1932). From McGhee (2007)

intersections of various states are the 18 boxes ( $3 \times 3 \times 2$ ) shown in the diagram—1IA, 1IIA, 2IB, and so on. Note, however, that only two regions of the morphospace are filled: sections 1AII and 2CI. We can also show our morphospace as in the top of Fig. 5, where one section was occupied but now is empty, or as in the bottom of Fig. 5, where one space is occupied and the other is about to be occupied.

Returning to the topic of developmental and functional constraints, we can model the two as shown in Fig. 6, where the solid line represents the functional-constraint boundary and the dotted line the developmental-constraint boundary. An evolving form must remain within the functional and developmentally possible regions of

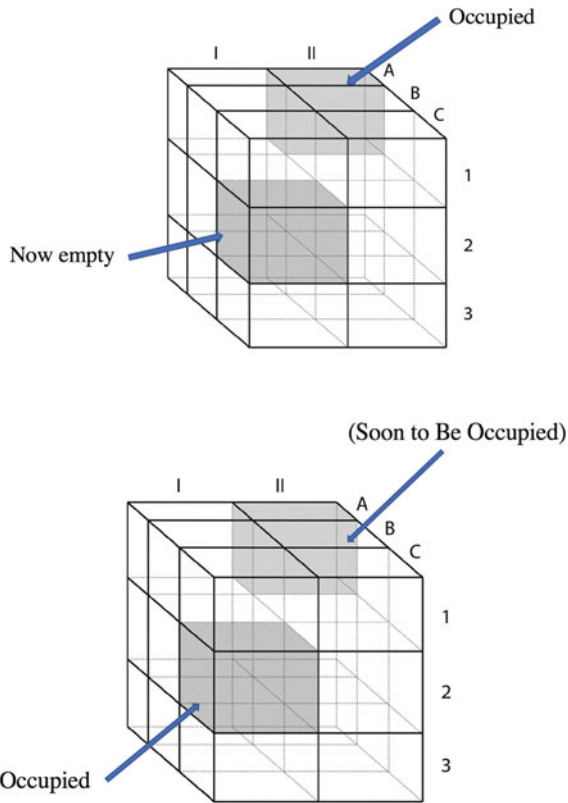


**Fig. 4** A simple theoretical morphospace defined by three traits—height, which has three possible states (1–3); depth which also has three (A–C); and width, which has two (I and II). Note that only two regions of the morphospace are filled: sections s1AII and 2CI

forms, although as shown in Fig. 6, the functional-constraint boundary does not have to spatially coincide with the developmental-constraint boundary. Mapping the functional and developmental-limit boundaries allows us to create a Venn diagram of four distinct sets of theoretical forms within our morphospace (McGhee 2011, 2018b): (1) forms that are both nonfunctional and developmentally impossible ( $f:0$ ); (2) forms that are both functional and can be developed ( $f:1$ ); (3) forms that can be developed but are nonfunctional ( $f:2$ ); and (4) forms that are functional but cannot be developed ( $f:3$ ). Charbonneau (2018) presents an excellent example of a form in  $f:3$ , a stone trilobate arrowhead—one that has three wings or blades. He asks why they are not found archaeologically and then provides an answer:

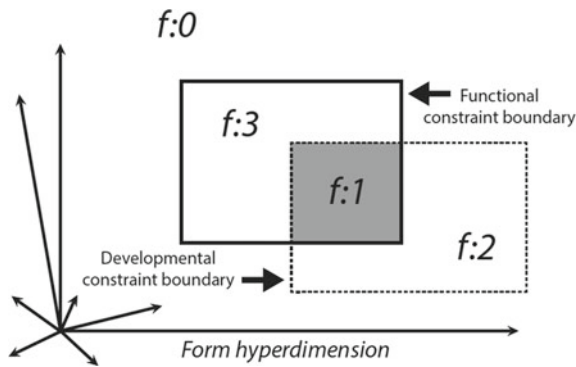
This has to do with the constraints imposed by the conchoidal fracturation process exploited by traditional flintknapping techniques. When a knapper produces such fractures on a core through percussion or pressure, the fissures travel roughly parallel to the surface of the core until they reach one of the core’s surfaces. Knapping trilobate arrowheads would necessitate that fractures stop somewhere halfway through the core and then come back toward the hammered platform’s surface, which contradicts the physical nature of the fracturing process. (p. 79)

Eren et al. (2018, p. 70) put it even more simply: “A knapper cannot strike a spherical flake. Nor can a knapper remove a cylindrical flake from the center of a core.” And indeed, trilobate arrowheads have been made from bone, ivory, and metal—but not from stone (Delrue 2007).



**Fig. 5** Same theoretical morphospace shown in Fig. 4, but now, in the upper diagram one of the two previously occupied sections is empty, and in the lower diagram one space is occupied and the other is about to be occupied

In a study separate from their experimental replication of Solutrean–Clovis overshoot flaking, Eren et al. (2018) examined the variability in stone-tool production flakes among experimentally knapped tools of different forms. The thinking was that if the shapes of production flakes generated from different reduction sequences substantially overlapped, this would signal a potentially important developmental constraint in stone-tool technology. This is because the result would be consistent with the idea that there exists only a limited variability in production-flake shape, and thus the probability for the parallel evolution of novelties is consequently higher than if production-flake shape was unbounded. Their results showed substantial overlap among the production-flake morphology of six different stone-tool reduction sequences. Although there were some statistical differences among the sets, there was far more similarity in terms of morphological variability. What was striking about the overlap was that not only did the experimental knapper (Eren) make tools of very different form, but the original stone nodule shapes and sizes were different



**Fig. 6** Spatial representations of functional and developmental constraint in theoretical morphospace. Boundaries of the solid-line rectangle delimit the functional-constraint boundary: forms located within this rectangle are functional, and forms outside the rectangle are nonfunctional. Boundaries of the dotted-line rectangle delimit the developmental-constraint boundary on possible form: forms within the dotted-line rectangle are developmentally possible, whereas forms outside the rectangle are developmentally impossible. Forms  $f:0$  are thus both nonfunctional and developmentally impossible; forms  $f:2$  are nonfunctional but developmentally possible; and forms  $f:3$  are functional but developmentally impossible. In contrast, forms  $f:1$  (gray-shaded region) are both functional and developmentally possible. From McGhee (2011)

and the knapper used different sets of tools to make the different forms. Despite all these sources of variability, there was still substantial overlap in production-flake shape.

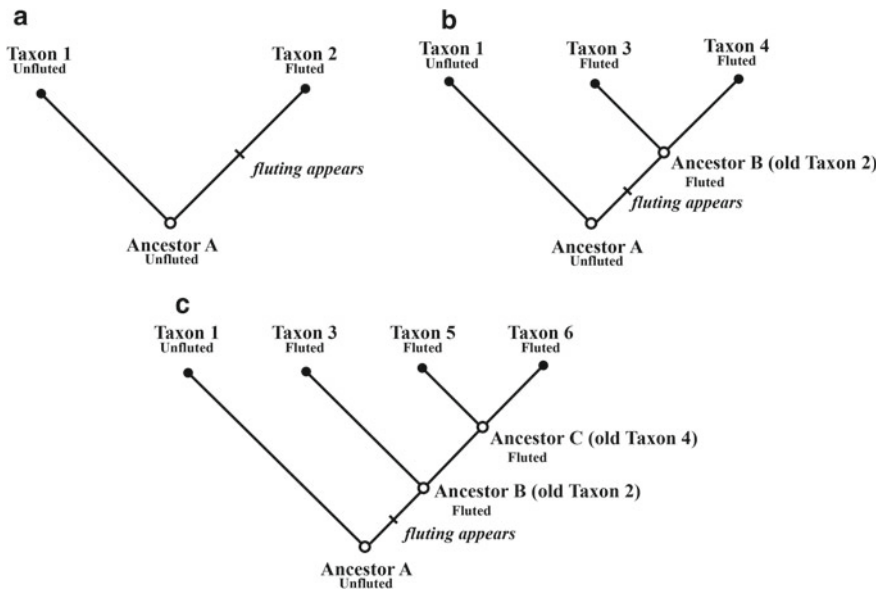
### 10.2.3 Phylogenetic Analysis

As important as experimental replication and the notion of theoretical morphospace are for understanding the available (and not available) pathways for reducing stone into functional tools, irrespective of time or place, we still need a means of ordering events not simply in terms of chronology but also in terms of phylogeny—what produced what. As we mentioned in Sect. 1, one phylogenetic method that is being used increasingly in archaeology is cladistics. The method defines phylogenetic relationships in terms of the relative recency of common ancestry: two groups, or taxa, are said to be more closely related to one another than either is to a third taxon if they share a common ancestor that is not also shared by the third taxon. The evidence for exclusive common ancestry is found in evolutionarily novel, or *derived*, traits.

The central tenet of cladistics is that not all phenotypic similarities are equally useful for reconstructing phylogenetic relationships. The method divides phenotypic similarities into three kinds. *Synapomorphies*, or shared derived traits, are similarities between two or more taxa that are inherited from the taxa's most recent common ancestor; *symplesiomorphies*, or shared ancestral traits, are similarities between two

or more taxa that are inherited from a more-distant common ancestor; and *homoplasies* are similarities that arise independently in two or more taxa—instances of convergence, irrespective of kind (discussed below). Of these three types of similarity, only synapomorphies are informative with regard to specific phylogenetic relationships.

How this might look with respect to stone projectile points is shown in Fig. 7. The phylogenetic trees show the evolution of a projectile-point lineage that begins with Ancestor A. For simplicity, we are tracking only a single trait, fluting—again, the removal of one or more longitudinal flakes from the base of a projectile point in order to thin it (Fig. 2). Over time, Ancestor A, which is unfluted, gives rise to two lines, one of which, like its ancestor, is unfluted and the other of which is fluted (Fig. 7a). Thus the trait “fluted” in Taxon 2 is derived from the ancestral trait, “unfluted.” In Fig. 7b, Ancestor B (old Taxon 2) gives rise to two new taxa, 3 and 4, each of which carries the derived trait, “fluted.” At this point “fluted” is a synapomorphy, or shared derived trait—one shared only by sister taxa and their immediate common



**Fig. 7** Phylogenetic trees showing the evolution of projectile-point taxa. In A, fluting appears during the evolution of Taxon 2 out of its ancestral group. Its appearance in Taxon 2 is as a derived trait. In B, Taxon 2 has produced two taxa, 3 and 4, both of which contain fluted specimens. The appearance of fluting in those sister taxa and their common ancestor makes it a synapomorphy (shared derived trait). In C, one of the taxa that appeared in the previous generation gives rise to two new taxa, 5 and 6, both of which contain fluted specimens. If we focus attention only on those two new taxa, fluting is now a symplesiomorphy (shared ancestral trait) because it is shared by more taxa than just sister taxa 5 and 6 and their immediate common ancestor. But if we include Taxon 3 in our focus, fluting is a synapomorphy because, following the definition, it occurs only in sister taxa 3, 5, and 6 and their immediate common ancestor. After O'Brien et al. (2001)

ancestor. In Fig. 7c, in which two descendent taxa have been added, fluting is now a symplesiomorphy—a shared ancestral trait—relative to taxa 5 and 6 because it is shared by three taxa and two ancestors. But relative to taxa 3, 5, and 6, fluting is a synapomorphy because it is shared only by three taxa and their immediate common ancestor, B. Thus depending on where in a lineage one begins, a trait can be derived or ancestral.

Whereas cladistics is designed to distinguish between phylogenetically informative traits and noninformative traits, other methods are not. Recall our earlier discussion of archaeologist David Clarke's (1968, p. 211) distinguishing between a "genetic and phyletic [phylogenetic] lineage or merely a phenetic and non-descent connected affinity." Phenetics—often referred to as "numerical taxonomy" (Sneath and Sokal 1973)—tells us only about overall similarity and nothing about historical relatedness. Whereas in phylogenetic analysis the evidence for exclusive common ancestry is the presence of evolutionarily novel, or derived, traits, phenetics places objects in groups according to the degree to which they are alike or not alike, with no distinction made among the kinds of traits used. It treats them all equally, irrespective of whether they are synapomorphies, symplesiomorphies, or homoplasies. Phenetics is bound to find similarity if it is present within a group of objects, but it neither establishes the existence of historical relationships nor demonstrates that the likeness indicates that sets of phenomena are related. Unfortunately, this is exactly the method Stanford and Bradley (2012) used as one of their bases for the Solutrean hypothesis.

### 10.2.3.1 Learning and Cultural Transmission

Phylogeny is created by the transmission of information, irrespective of mode. This means that cultural transmission—the process by which humans inherit, modify, and pass on information and behaviors—is as legitimate a mechanism for creating phylogenetic relationships as genetic transmission is (Collard and Shennan 2000; Collard et al. 2006; O'Brien et al. 2001, 2012; Platnick and Cameron 1977). Cultural transmission can be vertical in the sense of parent to offspring, analogous to genetic transmission, but it can also occur in the opposite direction—from offspring to parent. It can also be horizontal—between people of the same generation—as well as oblique—through unrelated people of different generations (Cavalli-Sforza and Feldman 1981).

*Learning* is the basis for cultural transmission because without it, there are no bits of information and behaviors to pass on. We can subdivide learning into two kinds—*individual* and *social*—keeping in mind that humans are neither purely social learners nor purely individual learners. Rather, certain conditions, perceived or real, dictate which one is used in any particular situation. Social learning is a particularly powerful adaptive strategy that allows others to risk failure so we don't have to (Henrich 2001; Laland 2004)—that is, it lets others filter behaviors and pass along those that have the highest payoff (Rendell et al. 2011). Social learning is how individuals learn their morals, language, technology, how to behave socially, and how to flake a Clovis

point. Over generations, the effect is cumulative, as individuals continue to “learn things from others, improve those things, transmit them to the next generation, where they are improved again, and so on” (Boyd and Richerson 2005, p. 4). Thus, learning and cultural transmission together create the means of heritability. Together they create traditions—persistent configurations in single technologies or other systems of related forms (Willey and Phillips 1958). Traditions are collections of related lineages, say, of projectile points, and both phenomena reflect transmission, persistence via replication, and heritable continuity (O'Brien et al. 2012)—the three legs of the evolutionary stool.

### 10.2.3.2 Arguments Against Cultural Phylogeny

Some anthropologists have argued that cultural phylogeny—what produced what—is nearly impossible to reconstruct because of the nature of cultural evolution (e.g., Bateman et al. 1990; Hornborg 2005; Terrell 2004), which they view as a different kind of process from biological evolution, with a faster tempo and often involving a different mode—horizontal transmission. This, they argue, creates reticulation—Darwin's (1859) “entangled bank”—which eradicates most or all traces of phylogenetic history, thus reducing the cultural landscape to little more than a blur of interrelated forms (O'Brien et al. 2013). This process is often referred to as *ethnogenesis*, defined broadly as cultural evolution that occurs “through the borrowing and blending of ideas and practices, and the trade and exchange of objects, among contemporary populations; the source of change is external” (Borgerhoff Mulder et al. 2006, p. 54). Cultural evolution probably is, in most respects, faster than biological evolution, and it can involve reticulation, but in our view these aspects are not necessarily problematic. For one thing, biological evolution can involve not only reticulation, where between-species hybridization might be as high as 15–25% in plants and as high as 10% in animals, but also cospeciation and lateral (horizontal) gene transfer (Gontier 2015).

Despite these issues, biologists have not thrown up their hands and abandoned the use of phylogenetic trees. Rather, they admit that the history of life is messy (Bell et al. 2010) and that there may, in fact, be no such thing as the “real” tree of life (see O'Malley et al. 2010)—or, if there is, we will never find it. Biologists and cultural phylogenists recognize that they deal with subtrees of that “tree” and that those subtrees are nothing more than models (Archibald et al. 2003; Collard et al. 2006). The key issue here is conflation of terms and concepts, especially *hybridization*, which has been used erroneously in cultural studies to denote any instance of horizontal transmission (e.g., Terrell et al. 1997). This equates process (hybridization) with mode (reticulation), which is specious.

Let us briefly consider units of three different scales: parental units, offspring units, and units of transmission. The mating of two parental organisms will produce an offspring with 50% of its genes originating with each parent—a 50/50 F1. Thus, the offspring is an even mixture of its parents in terms of the units of transmission. With respect to units of cultural transmission, horizontal transmission *might* produce

an offspring comprising equal parts of those replicators, but the odds are strong that it will not. To be an instance of hybridization, however, not only must something akin to a 50/50 F1 offspring be produced, but that hybrid must then transfer its mixture of genes into at least one of the parent species through *introgression*—gene flow from one species into the gene pool of another by repeated backcrossing of an interspecific hybrid with one of its parent species (Anderson 1949). Subsequent generations must next include the extralineage genes, and they must spread throughout the population in order to effect mongrelization (Levin 2002). If these extralineage genes spread in such a manner, then reticulation is the mode. If those extralineage genes do not spread in such a manner, then no hybrid mongrel species will be produced.

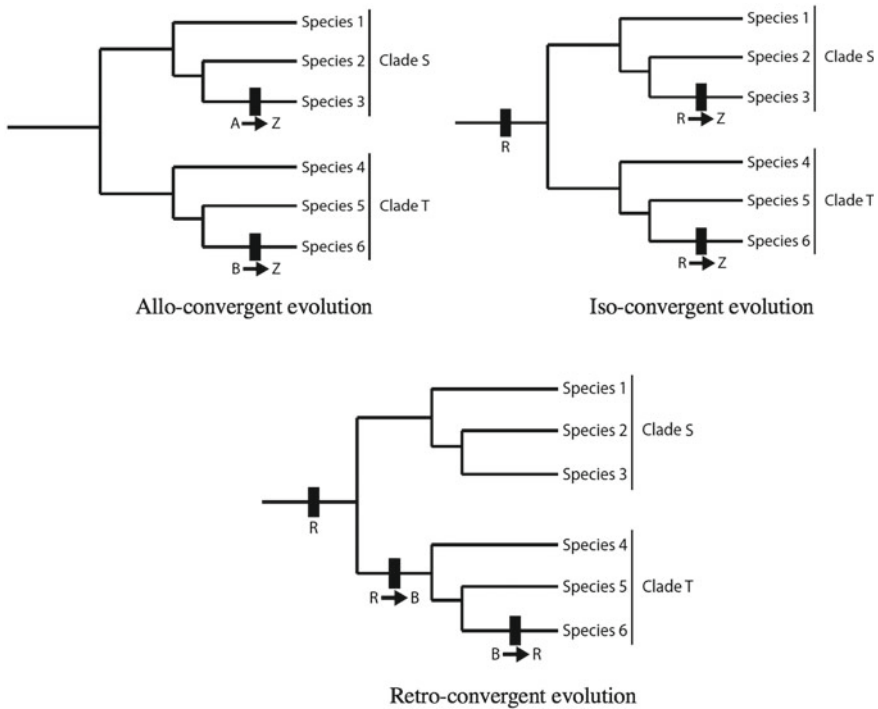
Goodenough (1997, p. 178) makes many of the same points with respect to language: “Contact between Japan and the United States has resulted in considerable borrowing in language and culture by Japan and some reverse borrowing by the United States, but their languages and cultures retain their respectively distinct phylogenetic identities.” Borrowing has not created a “hybrid” culture or language. Further, linguists do not flip a coin to determine whether two or more languages share a phylogenetic history. Numerous case studies have provided the basis for deciding which linguistic traits might be synapomorphies (shared derived traits) and which might be symplesiomorphies (Nichols 1996). Archaeologists interested in the evolution of stone tools use similar reasoning (e.g., Jennings and Smallwood 2018; Smallwood et al. 2018).

#### 10.2.4 *Convergence and Terminology*

All of these issues are imbedded in the concept of convergence and how it affects our ability to construct phylogenetic relations, whether among hominins, languages, or stone tools. With respect to homoplasy—a trait that arises independently in two or more taxa—suppose that the tree in Fig. 7 is a true depiction of projectile-point evolution. Further suppose that taxa 1 and 6 share a trait—say, beveling—that taxa 3 and 5 do not. We would refer to beveling as an instance of homoplasy. As straightforward as this sound, there are different mechanisms that can produce homoplasy, but our efforts to understand the evolutionary phenomenon of convergence have been hampered by a confusing terminology concerning the mechanisms and directionality of that phenomenon (McGhee 2019; McGhee et al. 2018; Pontarotti and Hue 2016). The terms include parallel evolution, reverse evolution, and convergent evolution in the strict sense of the word.

To overcome this terminological obstacle, McGhee et al. (2018; see also McGhee 2018a; Pontarotti and Hue 2016) specified three pathways by which evolution can produce convergence (Fig. 8). *Allo-convergent* evolution refers to the independent evolution of the same or very similar new trait from *different precursor traits* in different lineages; *iso-convergent* evolution refers to the independent evolution of the same or very similar new trait from the *same precursor trait* but in different lineages; and *retro-convergent* evolution refers to the independent *re-evolution* of the

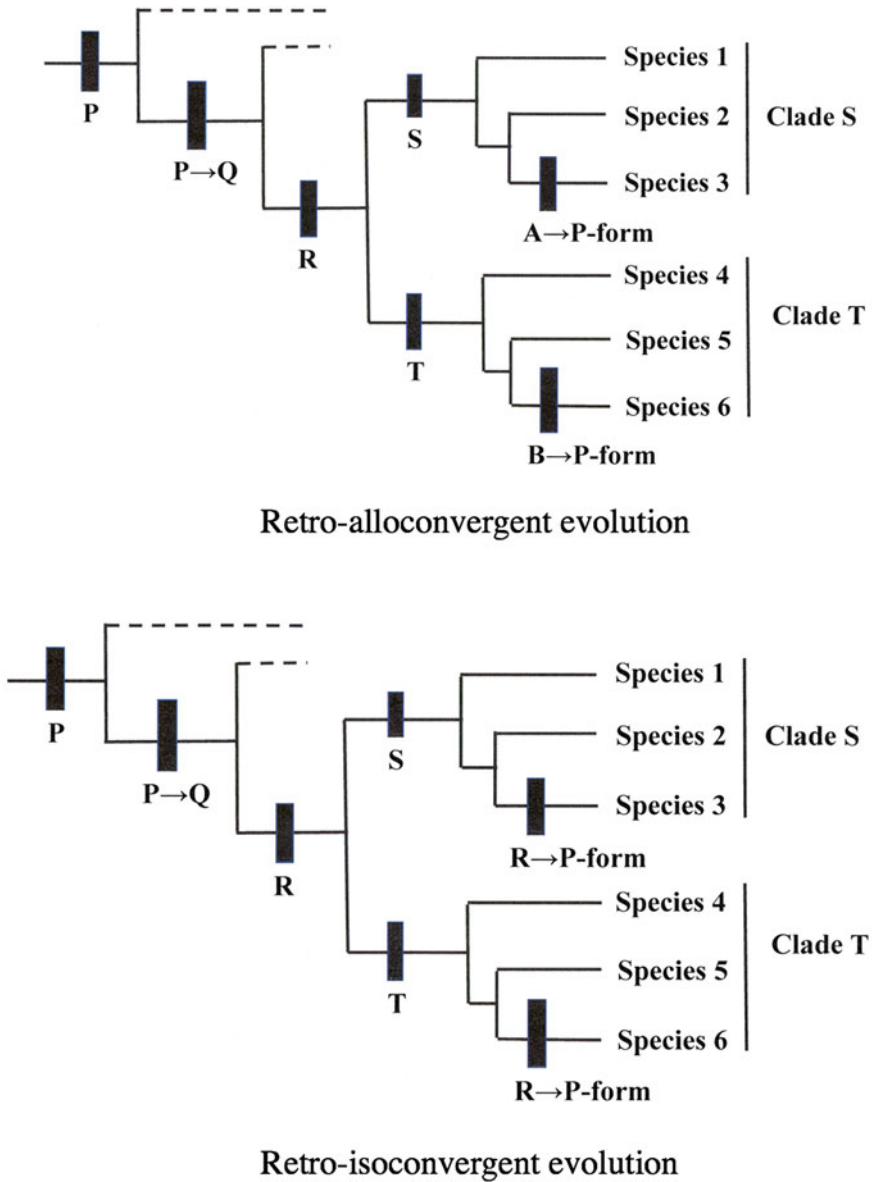




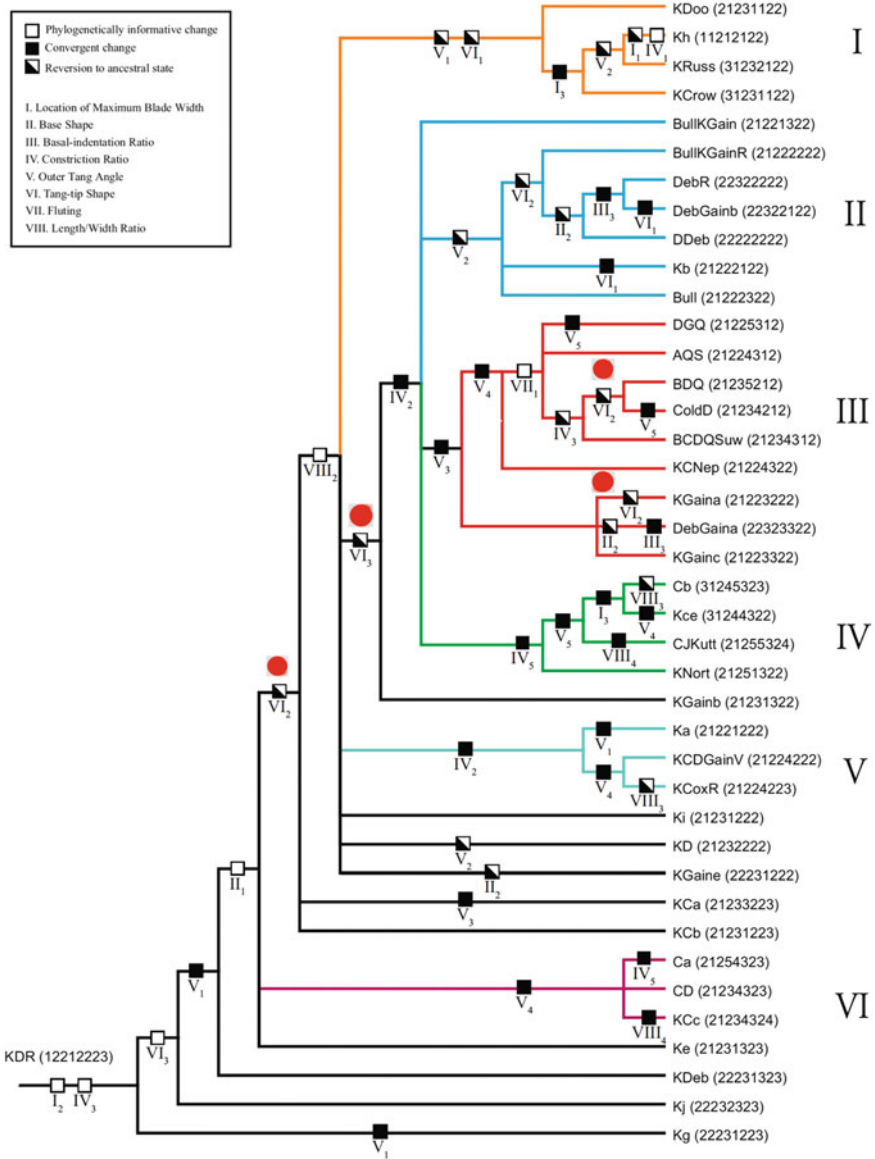
**Fig. 8** Hypothetical cladograms showing relations among six species in two clades. In allo-convergent evolution, two (or more) identical or very similar new traits (Z) are produced from different precursor traits in different lineages (A and B). In iso-convergent evolution, two (or more) identical or very similar new traits (Z) are produced from the same precursor trait (R) but in different lineages. In retro-convergent evolution, there is an independent re-evolution of the same or very similar trait to an ancestral trait in different lineages (R  $\rightarrow$  B then back to R). After McGhee (2019)

same or very similar trait to an *ancestral trait* in different lineages. Retro-convergent evolution itself contains two subtypes (Fig. 9): (1) *retro-alloconvergent* evolution, which refers to the independent *re-evolution* of the same or very similar trait to an ancestral trait from *different precursor traits* in different lineages; and (2) *retro-isoconvergent* evolution, which refers to the independent *re-evolution* of the same or very similar trait to an ancestral trait from the *same precursor trait* in different lineages (McGhee 2019).

This new terminology should find a comfortable home in archaeology. To date, the traits that have been used to create cultural phylogenies have been examined strictly from the standpoint of being phylogenetically informative or not. As an example, Fig. 10 shows a phylogenetic tree that one of us (MOB) constructed using 281 fluted projectile points (Clovis and related forms) from the eastern USA (O'Brien et al. 2014a, b). The tree contains 48 character-state changes, represented by squares. Shown in the box at the left are the eight characters that were used (Roman numerals). Each of the 48 squares is labeled with a Roman numeral indicating a character



**Fig. 9** Hypothetical cladograms (the same taxa and clades as in Fig. 8) illustrating two types of retro-convergent evolution. Here, ancestral trait P has been lost as a result of its evolutionary transition to a subsequent trait, Q. In retro-alloconvergence, a P-form trait very similar to ancestral trait P is re-evolved in species 3 and 6 from two different precursor traits, A and B. In retro-isoconvergence, a P-form trait very similar to ancestral trait P is re-evolved in species 3 and 6 from the same precursor trait, R. From McGhee (2019)



**Fig. 10** Phylogenetic tree of 41 classes of fluted projectile points from eastern North America, with six clades shown in different colors and demarcated by large Roman numerals. Small Roman numerals denote characters, and subscript numbers denote character states. Open squares indicate phylogenetically informative changes; shaded squares indicate parallel or convergent changes (homoplasy); and half-shaded squares indicate characters that reverted to an ancestral state. From O'Brien et al. (2015); courtesy Matt Boulanger

that has changed state; the subscript Arabic numeral indicates the evolved character state (see O'Brien et al. (2014a, b) for individual character states). White boxes indicate phylogenetically informative changes—shifts that result from descent with modification. Note that instances of homoplasy are labeled as being either cases of convergence (shaded squares), where knappers or groups of knappers landed on the same adaptive peaks through independent experimentation, or cases of reversion to an ancestral character state (half-shaded squares).

This terminology is not incorrect, but it overlooks important information that can be conveyed through use of the new terminology. For example, note character VI, which refers to the shape of the tang, or ear, tip (pointed, rounded, or blunt). Focusing just on classes in clade III (in red), note the two occurrences of character state VI<sub>2</sub> (marked by red circles), indicating the appearance of rounded tangs in different lineages but evolving from the same precursor state, VI<sub>3</sub> (blunt tangs) (red circle). This is a clear example of iso-convergent evolution. But let us go back a bit farther into the tree and see what other kinds of convergence might be present with respect to character VI. Note that the previous state of character VI is VI<sub>2</sub> (rounded) (red circle). This, then, makes the two occurrences of VI<sub>2</sub> in the red clade retro-isoconvergent and not simply iso-convergent.

In proposing their new terminology, McGhee et al. (2018) formulated a future research pathway that applies equally to archaeology as to biology and paleobiology: re-analyze cases of convergent evolution described in the literature, with the goal of determining whether (and if so, how) they share causative mechanisms that led to convergence. With respect to fluted points, for example, we know that various traits, especially those associated with the proximal end of a point—the area of attachment to a wooden foreshaft—constantly evolved and re-evolved. We need precise terminology in order to track the kinds of changes; to determine how the traits were, or were not, linked; to estimate rates of change; and to examine possible functional reasons for the changes. Evidence suggests that Late Pleistocene Clovis hunters continually modified their points to suit the characteristics of local prey and/or the habitats in which they hunted (Buchanan et al. 2014). As a result, there could have been tremendous differences in evolutionary rates among technological traits, which would represent a classic case of *mosaic evolution*, where unlinked traits assume different evolutionary histories (Carroll 1997).

### 10.3 Concluding Remarks

Scenarios such as the Solutrean hypothesis remind us about the potential dangers involved in taking less than a detailed methodological approach to distinguishing between archaeological instances of convergence and divergence (O'Brien et al. 2018). How many times throughout the history of archeology, one might ask, have instances of divergence been posited on a whole lot less evidence than what has been brought to bear in the Solutrean–Clovis debate? The answer must be somewhere in the thousands. As Foley and Lahr (2003, p. 110) put it, stone tools were “endlessly thrown

up convergently by the demands of the environment and social organization.” In other words, there are only so many ways to make stone tools, and unrelated toolmakers undoubtedly found common solutions to environmental “problems” countless times the world over, similar to what occurs in the natural realm, where convergence is now viewed as the dominant evolutionary process—the “evolutionary expectation rather than the exception” (McGhee 2019, p. 237).

How can archaeologists distinguish between convergence and divergence? We recommend an integrated approach involving experimental replication, cladistics, and the use of both precise terminology and key concepts such as morphospace and developmental and functional constraint. Cladistics is used to construct trees, which are hypothetical statements of relatedness. Those trees can then be used to examine not only the gross distribution of homoplasious traits across the various branches but the *kinds* of homoplasy. This use of trees to understand patterns of descent in order to then examine the distribution of adaptive (functional) features is the *modern comparative method*, which allows us to escape what Francis Galton pointed out in 1889: comparative studies of adaptation are irrelevant if we cannot rule out the possibility of a common origin of the adaptive features under examination (Naroll 1970). To escape Galton’s problem requires a working knowledge of the phylogeny of taxa included in an analysis. As Felsenstein (1985, p. 14) put it, “phylogenies are fundamental to comparative biology; there is no doing it without taking them into account.” The same applies to comparative studies of stone tools.

**Acknowledgements** We gratefully acknowledge the kind invitation of the organizers of the twenty-third Evolutionary Biology Meeting at Marseilles for the opportunity to present our work and to include it in this volume. In particular, we thank Pierre Pontarotti and Marie-Hélène Rome for their incredible kindness and hospitality, both during and after the meeting. We also thank an anonymous reviewer for extremely helpful comments on an early draft and Gloria O’Brien for error checking the manuscript.

## References

- Anderson E (1949) *Introgressive hybridization*. Wiley, New York
- Archibald JK, Mort ME, Crawford DJ (2003) Bayesian inference of phylogeny: a non-technical primer. *Taxon* 52:187–191
- Bateman R, Goddard I, O’Grady R, Funk VA, Mooi R, Kress WJ, Cannell P (1990) Speaking of forked tongues: the feasibility of reconciling human phylogeny and the history of language. *Curr Anthropol* 31:1–24
- Bell MA, Futuyma DJ, Eanes WF, Levinton JS (eds) (2010) *Evolution since Darwin: the first 150 years*. Sinauer, Sunderland, MA
- Bonner JT (1988) *The evolution of complexity*. Princeton University Press, Princeton, NJ
- Borgerhoff Mulder M, Nunn CL, Towner MC (2006) Cultural macroevolution and the transmission of traits. *Evol Anthropol* 15:52–64
- Boyd R, Richerson PJ (2005) *The origin and evolution of cultures*. Oxford University Press, Oxford
- Bradley B, Stanford D (2004) The North Atlantic ice-edge corridor: a possible Paleolithic route to the New World. *World Archaeol* 36:459–478

- Bradley BA, Collins MB, Hemmings A (2010) Clovis technology. *International Monographs in Prehistory*, Ann Arbor, MI
- Brew JO (1946) The archaeology of Alkali Ridge, southeastern Utah. Peabody Museum of Archaeology and Ethnology Papers, vol. 21. Harvard University, Cambridge, MA
- Buchanan B, O'Brien MJ, Collard M (2014) Continent-wide or region-specific? A geometric-morphometrics-based assessment of variation in Clovis point shape. *Archaeol Anthropol Sci* 6:145–162
- Carroll RL (1997) Patterns and processes of vertebrate evolution. Cambridge University Press, Cambridge
- Cavalli-Sforza LL, Feldman M (1981) Cultural transmission and evolution: a quantitative approach. Princeton University Press, Princeton, NJ
- Charbonneau M (2018) Technical constraints on the convergent evolution of technologies. In: O'Brien MJ, Buchanan B, Eren MI (eds) *Convergent evolution in stone-tool technology*. MIT Press, Cambridge, MA, pp 73–89
- Clarke DL (1968) *Analytical archaeology*. Methuen, London
- Collard M, Shennan SJ (2000) Ethnogenesis versus phylogenesis in prehistoric culture change: a case-study using European Neolithic pottery and biological phylogenetic techniques. In: Renfrew C, Boyle CK (eds) *Archaeogenetics: DNA and the population prehistory of Europe*. McDonald Institute for Archaeological Research, Cambridge, pp 89–97
- Collard M, Shennan SJ, Tehrani JJ (2006) Branching, blending, and the evolution of cultural similarities and differences among human populations. *Evol Hum Behav* 27:169–184
- Darwin C (1859) *On the origin of species by means of natural selection; or the preservation of favoured races in the struggle for life*. Murray, London
- Dawkins R (1982) *The extended phenotype: the long reach of the gene*. Oxford University Press, Oxford
- Delrue P (2007) Trilobate arrowheads at ed-Dur (U.A.E., Emirate of Umm al-Qaiwain). *Arab Archaeol Epigraphy* 18:239–250
- Eren MI, Buchanan B, O'Brien MJ (2018) Why convergence should be a potential hypothesis for the emergence and occurrence of stone-tool form and production processes: an illustration using replication. In: O'Brien MJ, Buchanan B, Eren MI (eds) *Convergent evolution in stone-tool technology*. MIT Press, Cambridge, MA, pp 61–71
- Eren MI, Lycett SJ, Patten RJ, Buchanan B, Pargeter J, O'Brien MJ (2016) Test, model, and method validation: the role of experimental stone artifact replication in hypothesis-driven archaeology. *Ethnoarchaeology* 8:103–136
- Eren MI, Patten RJ, O'Brien MJ, Meltzer (2013) Refuting the technological cornerstone of the Ice-Age Atlantic crossing hypothesis. *J Archaeol Sci* 40:2934–2941
- Eren MI, Patten RJ, O'Brien MJ, Meltzer DJ (2014) More on the rumor of “intentional overshot flaking” and the purported Ice-Age Atlantic crossing. *Lithic Technol* 39:55–63
- Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1–15
- Foley R (1987) Hominid species and stone-tool assemblages: how are they related? *Antiquity* 61:380–392
- Foley RA, Lahr MM (2003) On stony ground: lithic technology, human evolution, and the emergence of culture. *Evol Anthropol* 12:109–122
- Geribàs N, Mosquera M, Vergès JM (2010) What novice knappers have to learn to become expert stone toolmakers. *J Archaeol Sci* 37:2857–2870
- Gingerich JAM (2011) Down to seeds and stones: a new look at the subsistence remains from Shawnee-Minisink. *Am Antiquity* 76:127–144
- Gontier N (ed) (2015) *Reticulate evolution: symbiogenesis, lateral gene transfer, hybridization and infectious heredity*. Springer, Cham, Switzerland
- Goodenough WH (1997) Comment on “The dimensions of social life in the Pacific: human diversity and the myth of the primitive isolate” (Terrell JE, Hunt TL, Gosden C). *Curr Anthropol* 38:177–178
- Haynes G (2015) The millennium before Clovis. *PaleoAmerica* 1:134–162

- Hennig W (1966) *Phylogenetic systematics*. University of Illinois Press, Urbana
- Henrich J (2001) Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. *Am Anthropol* 103:992–1013
- Holden CJ, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. *Hum Biol* 69:605–628
- Hornborg A (2005) Ethnogenesis, regional integration, and ecology in prehistoric Amazonia. *Curr Anthropol* 46:589–620
- Jennings TA, Smallwood AM (2018) Clovis and Toyah: convergent blade technologies on the Southern Plains periphery of North America. In: O'Brien MJ, Buchanan B, Eren MI (eds) *Convergent evolution in stone-tool technology*. MIT Press, Cambridge, MA, pp 229–251
- Kelly RL (1994) Some thoughts on future directions in the study of stone tool organization. In: Carr P (ed) *The organization of North American prehistoric chipped stone tool technology*. International Monographs in Prehistory, no. 7, Ann Arbor, MI, pp 132–136
- Kroeber AL (1923) *Anthropology*. Harcourt, Brace, New York
- Kroeber AL (1931) Historical reconstruction of culture growths and organic evolution. *Am Anthropol* 33:149–156
- Kvitek DJ, Sherlock G (2011) Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet* 7(4):e1002056
- Laland KN (2004) Social learning strategies. *Learn Behav* 32:4–14
- Leonard RD (2001) Evolutionary archaeology. In: Hodder I (ed) *Archaeological theory today*. Polity, Cambridge, pp 65–97
- Leonard RD, Jones GT (1987) Elements of an inclusive evolutionary model for archaeology. *J Anthropol Archaeol* 6:199–219
- Levin DA (2002) Hybridization and extinction. *Am Sci* 90:254–261
- Lyman RL, O'Brien MJ (1998) The goals of evolutionary archaeology: history and explanation. *Curr Anthropol* 39:615–652
- Lyman RL, O'Brien MJ, Dunnell RC (1997) *The rise and fall of culture history*. Plenum, New York
- Mace R, Pagel M (1994) The comparative method in anthropology. *Curr Anthropol* 35:549–564
- McGhee GR (1999) *Theoretical morphology: the concept and its applications*. Columbia University Press, New York
- McGhee GR (2007) *The geometry of evolution: adaptive landscapes and theoretical morphospaces*. Cambridge University Press, Cambridge
- McGhee GR (2011) *Convergent evolution: limited forms most beautiful*. MIT Press, Cambridge, MA
- McGhee GR (2018a) Convergence. In: Nuño de la Rosa L, Müller G (eds) *Evolutionary developmental biology*. Springer, Cham, Switzerland. [https://doi.org/10.1007/978-3-319-33038-9\\_124-1](https://doi.org/10.1007/978-3-319-33038-9_124-1)
- McGhee GR (2018b) Limits on the possible forms of stone tools: a perspective from convergent biological evolution. In: O'Brien MJ, Buchanan B, Eren MI (eds) *Convergent evolution in stone-tool technology*. MIT Press, Cambridge, MA, pp 23–46
- McGhee GR (2019) *Convergent evolution on earth: lessons for the search for extraterrestrial life*. MIT Press, Cambridge, MA
- McGhee GR, Hue I, Dardaillon J, Pontarotti P (2018) A proposed terminology of convergent evolution. In: Pontarotti P (ed) *Origin and evolution of biodiversity*. Springer, Cham, Switzerland, pp 331–340
- Miller DS, Holliday VT, Bright J (2014) Clovis across the continent. In: Graf KE, Ketron CV, Waters MR (eds) *Paleoamerican odyssey*. Center for the Study of the First Americans, Texas A&M University, College Station, pp 541–560
- Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspina A-S, Sikora M, Reuther JD, Irish JD, Malhi RS, Orlando L, Song YS, Nielsen R, Meltzer DJ, Willerslev E (2018) Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* 553:203–207

- Naroll R (1970) Galton's problem. In: Naroll R, Cohen R (eds) *A handbook of method in cultural anthropology*. Columbia University Press, New York, pp 974–989
- Nichols J (1996) The comparative method as heuristic. In: Durie M, Ross M (eds) *The comparative method reviewed: regularity and irregularity in language change*. Oxford University Press, New York, pp 39–71
- O'Brien MJ (2019) More on Clovis learning: individual-level processes aggregate to form population-level patterns. *PaleoAmerica* 5:157–168
- O'Brien MJ, Bentley RA (2011) Stimulated variation and cascades: two processes in the evolution of complex technological systems. *J Archaeol Method Theory* 18:309–335
- O'Brien MJ, Boulanger MT, Buchanan B, Collard M, Lyman RL, Darwent J (2014) Innovation and cultural transmission in the American Paleolithic: phylogenetic analysis of eastern Paleoindian projectile-point classes. *J Anthropol Archaeol* 34:100–119
- O'Brien MJ, Boulanger MT, Buchanan B, Bentley RA, Lyman RL, Lipo CP, Madsen ME, Eren MI (2016) Design space and cultural transmission: case studies from Paleoindian eastern North America. *J Archaeol Method Theory* 23:692–740
- O'Brien MJ, Buchanan B, Collard M, Boulanger MT (2012) Cultural cladistics and the early prehistory of North America. In: Pontarotti P (ed) *Evolutionary biology: mechanisms and trends*. Springer-Verlag, Berlin, pp 23–42
- O'Brien MJ, Boulanger MT, Collard M, Buchanan B, Tarle L, Straus LG, Eren MI (2014) On thin ice: problems with Stanford and Bradley's Solutrean-Clovis hypothesis. *Antiquity* 88:606–624
- O'Brien MJ, Buchanan B, Boulanger MT, Mesoudi A, Collard M, Eren MI, Bentley RA, Lyman RL (2015) Transmission of cultural variants in the North American Paleolithic. In: Mesoudi A, Aoki K (eds) *Learning strategies and cultural evolution during the Palaeolithic*. Springer, Tokyo, pp 121–143
- O'Brien MJ, Buchanan B, Eren MI (2018) Issues in archaeological studies of convergence. In: O'Brien MJ, Buchanan B, Eren MI (eds) *Convergent evolution in stone-tool technology*. MIT Press, Cambridge, MA, pp 3–20
- O'Brien MJ, Collard M, Buchanan B, Boulanger MT (2013) Trees, thickets, or something in between? Recent theoretical and empirical work in cultural phylogeny. *Israel J Ecol Evol* 59:45–61
- O'Brien MJ, Darwent J, Lyman RL (2001) Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the southeastern United States. *J Archaeol Sci* 28:1115–1136
- O'Brien MJ, Lyman RL (2000) *Applying evolutionary archaeology: a systematic approach*. Kluwer Academic/Plenum, New York
- O'Brien MJ, Lyman RL (2003) *Cladistics and archaeology*. University of Utah Press, Salt Lake City
- O'Malley MA, Martin W, Dupré J (2010) The tree of life: introduction to an evolutionary debate. *Biol Philos* 25:441–453
- Odling-Smee FJ, Turner JS (2011) Niche construction theory and human architecture. *Biol Theory* 6:283–289
- Platnick NI, Cameron D (1977) Cladistic methods in textual, linguistic, and phylogenetic analysis. *Sys Zool* 26:380–385
- Pontarotti P, Hue I (2016) Road map to study convergent evolution: a proposition for evolutionary systems biology approaches. In: Pontarotti P (ed) *Evolutionary biology*. Springer, Cham, Switzerland, pp 3–21
- Rendell L, Boyd R, Enquist M, Feldman MW, Fogarty L, Laland KN (2011) How copying affects the amount, evenness and persistence of cultural knowledge: insights from the Social Learning Strategies Tournament. *Phil Trans R Soc B* 366:1118–1128
- Rowe JH (1966) Diffusionism and archaeology. *Am Antiquity* 31:334–337
- Simpson GG (1961) *Principles of animal taxonomy*. Columbia University Press, New York
- Smallwood AM, Smith HL, Pevny CD, Jennings JA (2018) The convergent evolution of serrated points on the Southern Plains-Woodland border of Central North America. In: O'Brien MJ,



- Buchanan B, Eren MI (eds) *Convergent evolution in stone-tool technology*. MIT Press, Cambridge, MA, pp 203–227
- Sneath PHA, Sokal RR (1973) *Numerical taxonomy*. Freeman, San Francisco
- Stanford D, Bradley B (2002) Ocean trails and prairie paths? Thoughts about Clovis origins. In: Jablonski NG (ed) *The first Americans: the Pleistocene colonization of the New World*. Memoirs, vol 27. California Academy of Sciences, San Francisco, pp 255–271
- Stanford D, Bradley B (2012) *Across Atlantic ice: the origin of America's Clovis culture*. University of California Press, Berkeley
- Steward JH (1941) Review of "Prehistoric culture units and their relationships in northern Arizona" (Colton HS). *Am Antiquity* 6:366–367
- Story BA, Eren MI, Thomas K, Buchanan B, Meltzer DJ (2019) Why are Clovis fluted points more resilient than non-fluted lanceolate points? A quantitative assessment of breakage patterns between experimental models. *Archaeometry* 61:1–13
- Stout D (2011) Stone toolmaking and the evolution of human culture and cognition. *Phil Trans R Soc B* 366:1050–1059
- Straus LG (2000) Solutrean settlement of North America? A review of reality. *Am Antiquity* 65:219–226
- Straus LG (2005) The Upper Paleolithic of Cantabrian Spain. *Evol Anthropol* 14:145–158
- Terrell JE (2004) Review of "Cladistics and archaeology" (O'Brien MJ, Lyman RL). *J Anthropol Res* 60:303–305
- Terrell JE, Hunt TL, Gosden C (1997) The dimensions of social life in the Pacific: human diversity and the myth of the primitive isolate. *Curr Anthropol* 38:155–195
- Thomas DH (1986) Points on points: a reply to Flenniken and Raymond. *Am Antiquity* 51:619–627
- Thomas KA, Story BA, Eren MI, Buchanan B, Andrews BN, O'Brien MJ, Meltzer DJ (2017) Explaining the origin of fluting in North American Pleistocene weaponry. *J Archaeol Sci* 81:23–30
- Turner JS (2000) *The extended organism: the physiology of animal-built structures*. Harvard University Press, Cambridge, MA
- Turner JS (2012) Evolutionary architecture? Some perspectives from biological design. *Archit Design* 82:28–33
- Willey GR (1953) Archaeological theories and interpretation: New World. In: Kroeber AL (ed) *Anthropology today*. University of Chicago Press, Chicago, pp 361–385
- Willey GR, Phillips P (1958) *Method and theory in American archaeology*. University of Chicago Press, Chicago
- Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones DF (ed) *Proceedings of the Sixth Congress on Genetics*, vol 1. Brooklyn Botanic Garden, New York, pp 356–366
- Wright S (1988) Surfaces of selective value revisited. *Am Nat* 131:115–123

# Chapter 11

## Diversity and Evolution of RNase P



Isabell Schencking, Walter Rossmannith, and Roland K. Hartmann

**Abstract** Ribonuclease P (RNase P) is the essential endonuclease responsible for the 5'-end maturation of tRNAs. It is found in all forms of life, yet in an unprecedented variety of architectures. RNase P enzymes are, on the one hand, represented by RNA-based forms, where a structurally conserved, catalytic RNA molecule is associated with one or more (up to ten different) proteins. The ancient RNA apparently independently recruited protein(s) in the bacterial and archaeal/eukaryal lineages, and the protein moiety increased in number and mass in the latter, obviously at the expense of RNA's structural autonomy. Protein-only enzymes, representing the other principal form of RNase P, arose independently twice in evolution. In a few bacteria, a small protein (called HARP) has replaced the RNA enzyme; the protein is also found in some archaea, where it, curiously enough, coexists with an RNA-based RNase P. A distinct form of protein-only RNase P (PRORP) apparently originated at the root of eukaryal evolution. In lineages of four of the five eukaryal supergroups, PRORP replaced the RNA-based enzyme in one or more of the cellular compartments that harbor a tRNA processing machinery. In metazoan mitochondria, PRORP became dependent on two other mitochondrial enzymes in a multi-enzyme assembly. In addition to introducing the reader into the history of RNase P discoveries and the various enzyme architectures, including their phylogenetic distribution, we discuss the evolution of the RNase P-enzyme family in terms of origin, principles, and mechanistic scenarios.

---

I. Schencking · R. K. Hartmann (✉)

Institute of Pharmaceutical Chemistry, Philipps-University Marburg, Marbacher Weg 6,  
35037 Marburg, Germany

e-mail: [roland.hartmann@staff.uni-marburg.de](mailto:roland.hartmann@staff.uni-marburg.de)

W. Rossmannith

Center for Anatomy and Cell Biology, Medical University of Vienna, Währinger Straße 13,  
1090 Vienna, Austria

© Springer Nature Switzerland AG 2020

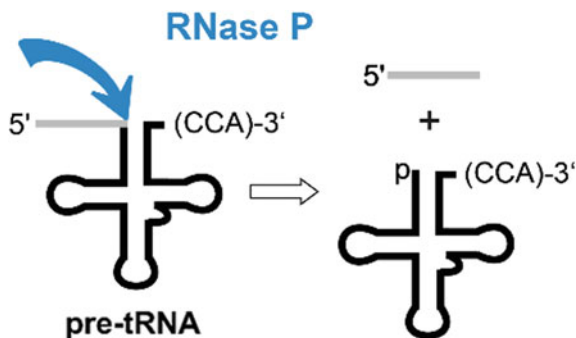
P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_11](https://doi.org/10.1007/978-3-030-57246-4_11)

255

## 11.1 Ribonuclease P: A Historical Introduction

Transfer RNAs (tRNAs) are generally transcribed with extra sequences at their 5'- and 3'-ends, and in some cases intervening sequences (introns). For their function as adaptor molecules in protein synthesis, tRNA primary transcripts have to undergo a series of post-transcriptional modifications, including endonucleolytic removal of 5'-precursor sequences by RNase P (Fig. 11.1), 3'-terminal trimming by 3'-exonucleases or, alternatively, endonucleolytic cleavage at the discriminator nucleotide by RNase Z, CCA addition to the 3'-end by an ATP(CTP):tRNA nucleotidyltransferase (CCA-adding enzyme), if present removal of introns by a specific tRNA-splicing machinery, and finally the introduction of a high number of diverse nucleoside modifications. Cleavage by RNase P is usually a first or early step of post-transcriptional tRNA maturation.

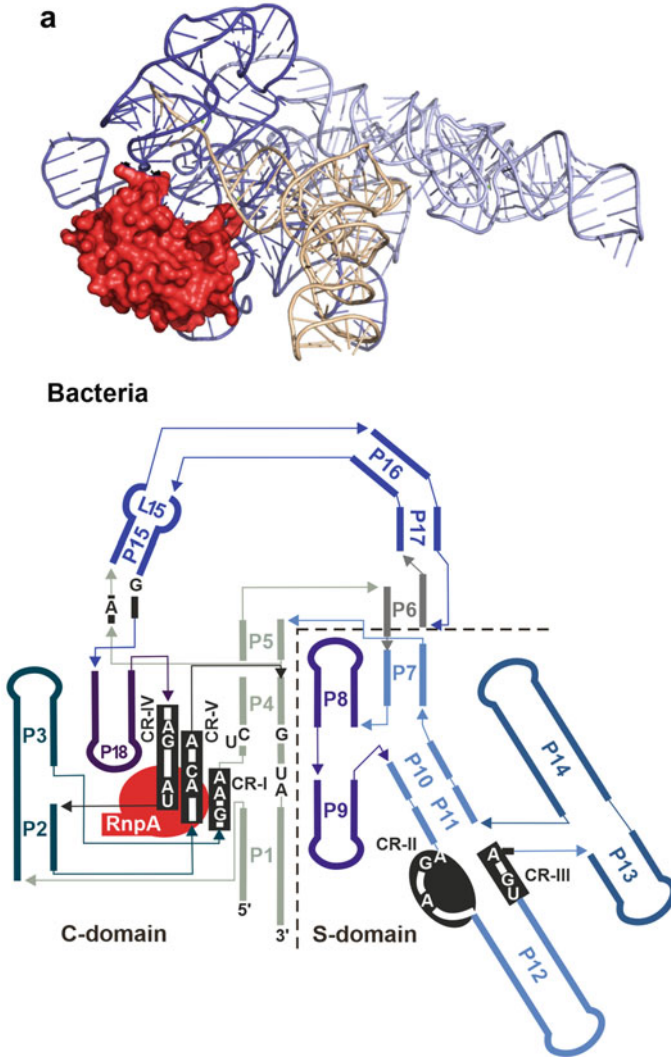
Around 1980, the laboratories of Sidney Altman and Norman R. Pace discovered the RNA component of bacterial RNase P (encoded by the *rnpB* gene) as the catalytic subunit of the ribonucleoprotein (RNP) enzyme (Guerrier-Takada et al. 1983). Nevertheless, the protein, termed RnpA, P protein or C5 (a specific abbreviation for *Escherichia coli* RnpA), is essential for enzyme function under in vivo conditions (Gösringer and Hartmann 2007). After its discovery, RNase P was long thought to consist of a catalytic RNA (ribozyme) subunit and one or more protein cofactors in all domains of life. This view was reinforced upon discovery of archaeal RNase P RNAs (Nieuwlandt et al. 1991; LaGrandeur et al. 1993; Haas et al. 1996), some of which showed RNA-alone activity in vitro as well (Pannucci et al. 1999), and finally by the observation that even the RNA component of human nuclear RNase P, although forming a holoenzyme with ten protein cofactors, retained low residual



**Fig. 11.1** Schematic depiction of tRNA 5'-end maturation by RNase P. The specific endonucleolytic cleavage is indicated by the blue arrow, which separates the 5'-leader (in gray) of precursor tRNAs from the body of the tRNA. The phosphodiester hydrolysis reaction releases the 5'-mature tRNA with a 5'-phosphate and the 5'-leader with a 3'-OH group. The CCA terminus is shown in parentheses to indicate that cleavage by bacterial RNases P is in most cases supported by the presence of 3'-CCA (except for cyanobacteria; Pascual and Vioque 1999), but does not play a role for processing by archaeal type M and eukaryotic nuclear RNase P (reviewed in Klemm et al. 2016)

yet specific RNA-alone activity *in vitro* (Kikovska et al. 2007). However, doubts regarding this universal RNA-centric nature of RNase P arose in the course of studies on plant and human mitochondrial RNase P. Peter Gegenheimer and coworkers reported in 1988 (Wang et al. 1988) that the spinach chloroplast RNase P activity showed a lower, protein-typical density, far below that of RNA-rich RNA:protein complexes, such as *E. coli* RNase P. Likewise, one of us reported in 1995 that human nuclear and mitochondrial RNase P activities are distinct enzymes based on their substrate specificities and physicochemical properties (Rossmanith et al. 1995), with the mitochondrial activity having the properties of a protein rather than an RNA:protein complex (Rossmanith and Karwan 1998). These findings sparked a smoldering and decade-long dispute that was long dominated by the RNA-centric majority in this research field, one reason being that the genes for protein-only RNase P enzymes could not be identified at that time. However, in 2008, one of us succeeded in disclosing the composition of human mitochondrial RNase P as a complex of three different proteins lacking any RNA component (Holzmann et al. 2008). One of the three protein subunits (PRORP/MRPP3) was identified as the endonuclease activity and bioinformatic searches then identified homologs in many eukarya, including land plants like *Arabidopsis thaliana* or protists like *Trypanosoma brucei*. These homologs were subsequently shown to act as single-polypeptide RNases P in nuclei and mitochondria/chloroplasts and were termed PRORP for proteinaceous or protein-only RNase P (Gobert et al. 2010; Gutmann et al. 2012; Taschner et al. 2012). The size of PRORPs (~60 kDa) is well in the range of the ~70 kDa inferred by Peter Gegenheimer and coworkers for spinach chloroplast RNase P based on size-exclusion chromatography (Thomas et al. 1995), so in retrospect, his group was obviously close to PRORP discovery.

The uniqueness of RNase P lies in the enzyme's unprecedented architectural diversity. The ancient form of RNase P consists of an RNA molecule of common ancestry that assembles into a complex with a variable number of protein subunits (1 in Bacteria, 5 in Archaea, and up to 10 in Eukarya), where the single RNA component represents the catalytic core of the enzyme (Fig. 11.2). Another form of RNase P, PRORP, is of independent origin and lacks any RNA subunit. It is found in various eukarya, but not in bacteria or archaea. Bioinformatic searches have identified PRORP homologs in four of the five eukaryal supergroups (Lechner et al. 2015), where they may act in the nucleus and/or in organelles. *A. thaliana*, for example, encodes three PRORP isoenzymes: PRORP1 localizes to mitochondria and chloroplasts, while PRORP2 and 3 function in the nucleus (Gobert et al. 2010; Gutmann et al. 2012). While the existence of protein-only RNases P in Eukarya became firmly accepted by 2008, the exclusive presence of RNA-based RNase P in Bacteria and Archaea remained cast in stone in the years that followed. However, in 2017, our groups published the discovery of yet another, unrelated protein-only RNase P in the hyperthermophilic bacterium *Aquifex aeolicus* and close relatives of the family Aquificaceae (Nickel et al. 2017). Bioinformatic analyses identified homologs of Aquifex RNase P (HARPs) in a few other bacteria, yet in many archaea belonging to the Euryarchaeota (Nickel et al. 2017; Daniels et al. 2019). Interestingly, HARPs and the nuclease domain of PRORPs belong to the same PIN domain-like supergroup



**Fig. 11.2** RNA-based holoenzymes. RNase P structures with tRNA bound representing all three domains of life, illustrated as 3D (top) and 2D (bottom) structures. **a** *T. maritima* RNase P representing type A bacterial RNase P (PDB: 3Q1Q). **b** Structure of *M. jannaschii* RNase P representing type M archaeal RNase P (PDB: 6K0B). **c** Structure of human RNase P:tRNA complex (PDB: 6AHU). In the 3D structure representations, the C-domain is colored in dark blue (only visible in panels a and b), the S-domain in light blue, and the tRNA is shown in sand color. The color code for the protein subunits is comparable in the corresponding 2D and 3D presentations. In the 2D models, coaxially stacked helices of the P RNA are shown in the same color and the dashed line separates C-domain and S-domain. The conserved regions (CRs; see Fig. 11.3c for details) are highlighted in black. The protein spheres are not drawn to scale. Homologous proteins in archaeal and eukaryal RNase P share the same color and proteins are positioned in areas where they interact with the P RNA according to the atomic structures

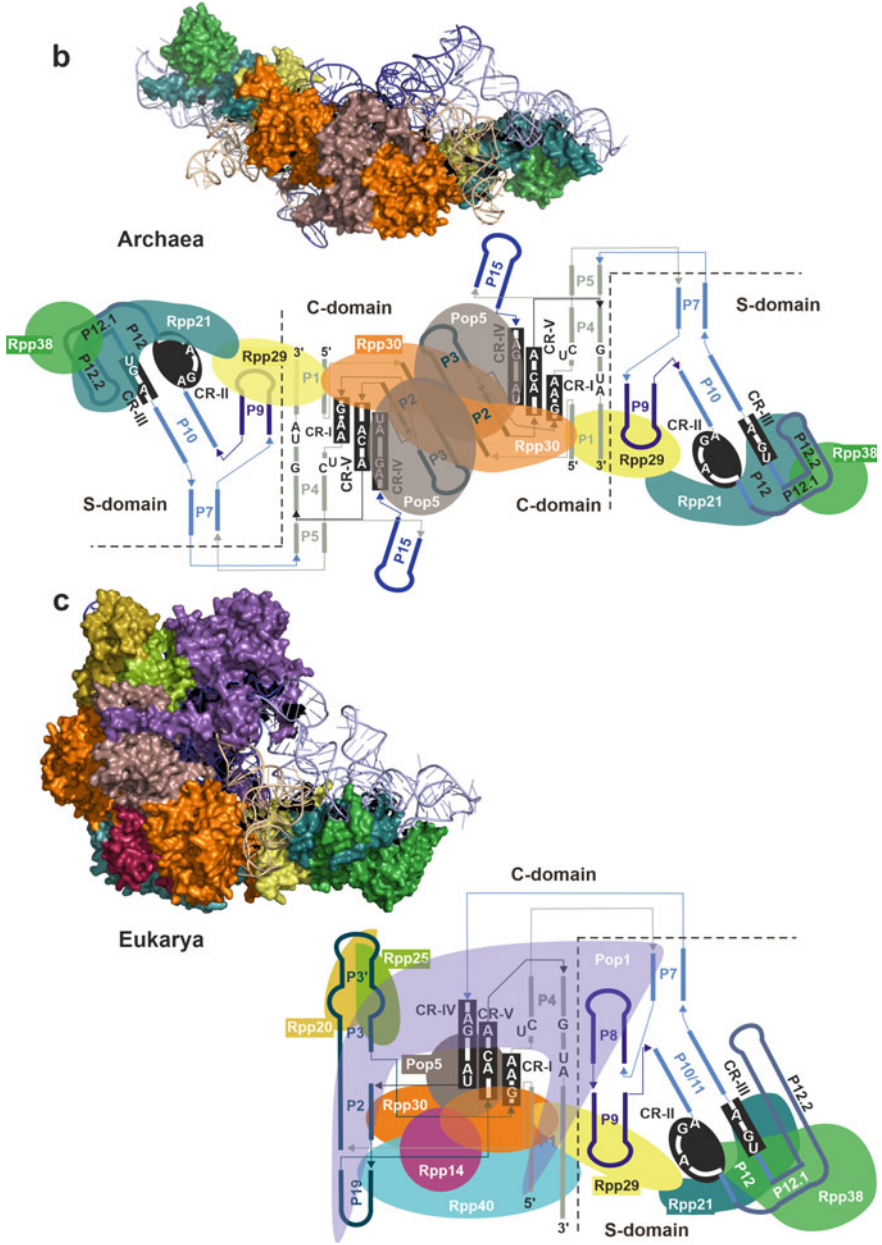


Fig. 11.2 (continued)

of endonucleases, although they are assigned to different subgroups (Matelska et al. 2017; Gobert et al. 2019).

Given the extreme diversity and independent origin of RNase P forms, RNase P can probably no more be considered a universally conserved enzyme, yet it still remains a universally conserved enzymatic activity, although a so far singular exception even challenges this latter rule; the parasitic archaeon *Nanoarchaeum equitans* manages to transcribe leaderless tRNAs whereby it finally obviated the need for RNase P activity (Randau et al. 2008).

## 11.2 A Panoply of Ribonucleoprotein Enzymes Based on a Conserved Catalytic RNA Found Throughout All Domains of Life

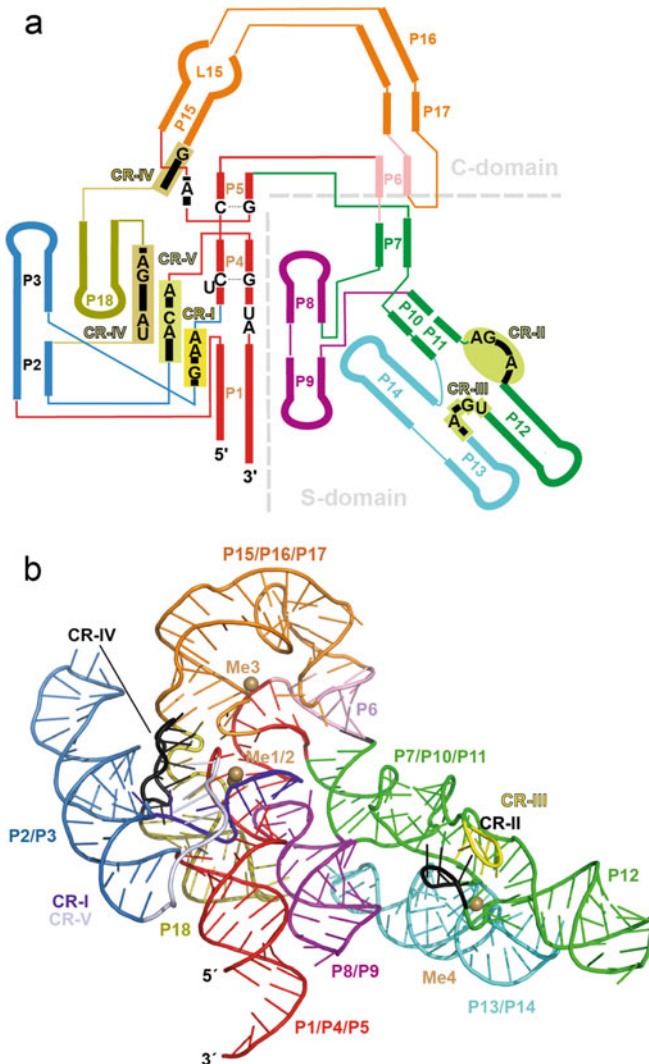
### 11.2.1 Bacterial RNase P: The Closest to the Ancestral RNA Enzyme

The catalytic RNA subunit of bacterial RNase P (P RNA), ~400 nt in size, folds into a compact structure with the help of  $Mg^{2+}$  ions (Fig. 11.3a, b). The RNA consists of two domains: the catalytic domain (C-domain) interacts with the cleavage site, the 5'-terminal portion of the acceptor stem and the tRNA 3'-CCA end; the specificity domain (S-domain) contacts the corner of the tRNA L-shape where D-loop and T-loop interact with each other (Fig. 11.4a, b). The conserved catalytic core is composed of a subset of conserved helices and single-stranded junctions that are required for catalysis by a two-metal-ion mechanism (Steitz and Steitz 1993; Warnecke et al. 1996). Five conserved regions (CR) I–V are shared among all bacterial P RNAs (Fig. 11.3c) and are still identifiable in archaeal and eukaryal P RNAs (Chen and Pace 1997) (Figs. 11.2 and 11.5). Beyond these conserved elements, peripheral helix/hairpin elements that show some variability among bacteria, stabilize the RNA's global structure by forming short-range and long-range docking interactions (Fig. 11.5a). Based on those variable helical elements, bacterial P RNAs are classified into two major architectural types: the most common A-type (ancestral), represented by, e.g., *E. coli* and *Thermotoga maritima*, and type B (*Bacillus*) mostly found in low GC gram-positive bacteria and typified by *Bacillus subtilis* (Fig. 11.5a). A third type C (*Chloroflexi*) architecture, present in *Chloroflexi*, is intermediate between type A and type B (Haas and Brown 1998). The single protein cofactor, although contributing as little as ~10% to the molar mass of the holoenzyme, is essential *in vivo*; yet *in vitro*, at elevated  $Mg^{2+}$  concentrations, the bacterial P RNA can catalyze the specific endonucleolytic tRNA 5'-end maturation reaction in the absence of the P protein (Guerrier-Takada et al. 1983). The protein binds to the C-domain of the RNA, close to the active site (Fig. 11.4a). The so-called RNR-motif, a basic amino acid sequence motif within the most conserved region of the protein, is crucial for this interaction. Although P proteins share only little

sequence identity beyond this conserved region, their overall structure is very similar (Kazantsev et al. 2003; Spitzfaden et al. 2000; Stams et al. 1998). Accordingly, bacterial P proteins and P RNAs are functionally interchangeable between different bacteria (Guerrier-Takada et al. 1983), as confirmed by genetic complementation experiments (Gösringer and Hartmann 2007; Waugh and Pace 1990; Wegscheid et al. 2006). P proteins possess a central cleft which exclusively binds the single-stranded 5'-leader of precursor tRNAs. Thereby, the protein contributes to catalytic efficiency without direct participation of amino acid residues in catalysis. Several partial contributions to holoenzyme function have been described for P proteins: By binding to the 5'-leader of substrate RNAs, the protein largely increases enzyme-substrate affinity under physiological salt conditions, where 5'-mature tRNAs bind with low affinity to RNase P (Crary et al. 1998; Kurz et al. 1998; Niranjanakumari et al. 1998). This enhances product (mature tRNA) release, a step that limits P RNA-alone reactions under multiple-turnover conditions in vitro (Reich et al. 1988; Tallsjö and Kirsebom 1993) and prevents product inhibition in vivo, where mature tRNA concentrations largely exceed those of 5'-precursor tRNAs at limiting RNase P concentrations (reviewed in Gößringer et al. 2020). By aligning and fixating the 5'-leader in the active site, as well as through inducing local conformational changes upon binding to P RNA which stabilizes the RNA's catalytic core (Buck et al. 2005; Guo et al. 2006), the P protein further enhances the affinity of functionally important  $Mg^{2+}$  ions in enzyme-substrate complexes to enhance the catalytic efficiency at physiological  $Mg^{2+}$  concentrations (Kurz and Fierke 2002; Sun and Harris 2007). The protein also equalizes the affinities of the structurally diverse precursor tRNAs within the cellular tRNA pool (Sun et al. 2006), and last but not least alleviates electrostatic repulsion effects between backbone phosphates of P RNA and precursor tRNA, the first assigned function of the P protein (Reich et al. 1988). Based on numerous biochemical studies, a mechanistic model of the reaction catalyzed by the bacterial RNase P holoenzyme was proposed: initially, the enzyme binds precursor tRNA (pre-tRNA) in a first bimolecular collision step near the diffusion limit and mainly recognizes the tRNA moiety in this initial complex. A second unimolecular conformational step, optimally requiring a 5'-leader length of  $\geq 4$  nt, increases the overall substrate affinity and is achieved by a decrease in the rate constant for the reversal of the conformational step. This isomerization step, coupled to interactions of the 5'-leader and the P protein, adjusts the position of essential functional groups including catalytic metal ions for efficient catalysis (Hsieh and Fierke 2009). Finally, the bacterial holoenzyme acts more efficiently on non-tRNA substrates than the RNA subunit alone (reviewed in Hartmann et al. 2009). Therefore, the protein subunit also contributes to broadening the substrate spectrum of the bacterial RNase P enzyme. Natural non-tRNA substrates of bacterial RNases P are hairpin RNAs that mimic the tRNA acceptor stem and T arm, including precursors to the signal recognition particle RNA (4.5S RNA) in  $\gamma$ -proteobacteria and tmRNA (aka 10Sa RNA). Other non-tRNA substrates of RNase P identified in *E. coli* comprise phage-induced regulatory RNAs, such as the phi80-induced M3 RNA, CI RNA of satellite phage P4 or C4 repressor RNA of bacteriophages P1/P7 (Bothwell et al. 1976; Forti et al. 1995; Hartmann et al. 1995). Studies demonstrating specific processing of small



RNA duplex substrates and even single-stranded RNA oligonucleotides, as well as reports on RNase P cleavages in riboswitches and intergenic regions of polycistronic transcripts in *E. coli*, *B. subtilis* and *Salmonella typhimurium* may suggest that the enzyme has a broader substrate spectrum than anticipated (reviewed in Hartmann et al. 2009). Yet, the metabolic importance of these endonucleolytic cleavages in mRNA transcripts needs to be examined.



**Fig. 11.3** Structure of bacterial RNase P RNA. **a, b** Juxtaposition of *T. maritima* P RNA secondary (a) and tertiary (b) structure (adapted from Reiter et al. 2010). **a** Coaxially stacked helices are depicted in the same color and the dashed gray line demarcates the C-domain and S-domain. The five conserved regions (CR-I to CR-V) are highlighted in yellow/lime green/ocher tones, with some conserved base identities indicated by bold letters (for more details, see panel c). **b** Tertiary structure

of *T. maritima* P RNA. Structural elements are colored as in panel a, while different colors were chosen for highlighting the conserved regions. Sand-colored spheres display the two active site metal ions (Me1 and Me2) as well as two structurally important metal ions (Me3 and Me4). The view visualizes the following features: (1) the two clusters of CR regions, CR-I/IV/V on the one hand and CR-II/III on the other hand, (2) the high-affinity metal ion-binding site (Me4) that supports the structural organization of CR-II/III, (3) the position of P8/P9 sandwiched between C-domain and the more distal parts of the S-domain, and (4) the P15/16/17/6 round arch into which the tRNA 3'-NCCA end is threaded with support from Me3. **c** Bacterial minimum consensus structure according to Siegel et al. (1996) and Marquez et al. (2005), adapted to the secondary structure presentation shown in panel a. The CR regions are highlighted in the same color code as in panel a; note that CR-I and CR-V extend into helix P4, which is not evident in panel a. Gray elements, which are evolutionarily volatile, are not part of the minimum consensus structure; they are shown for comparability with the *T. maritima* P RNA structure in panel a

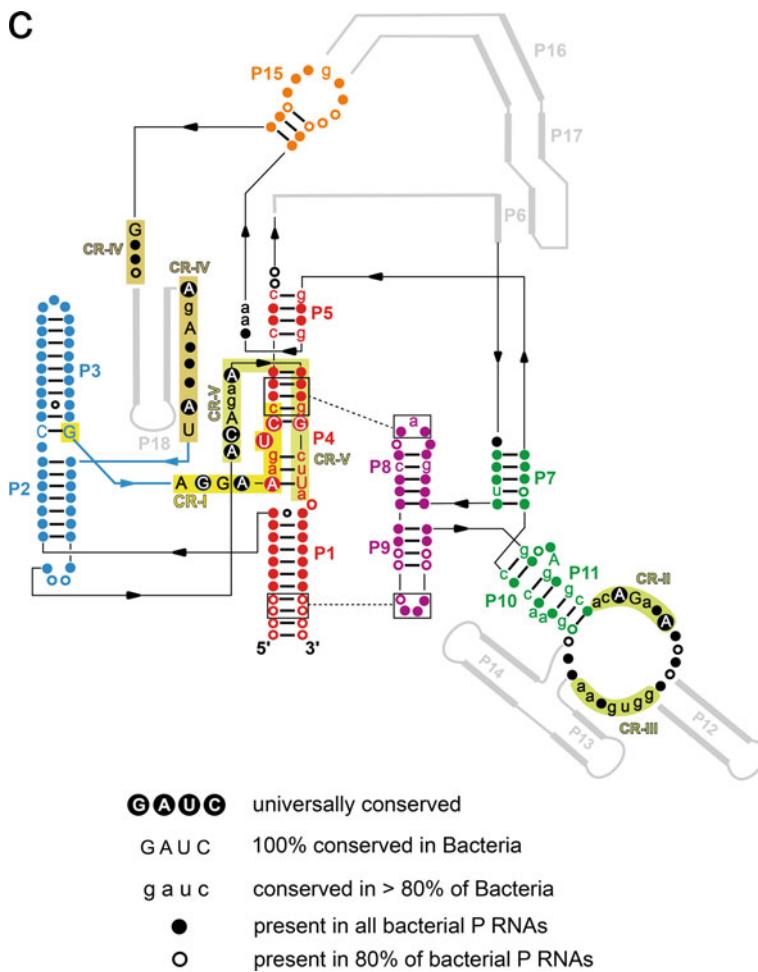
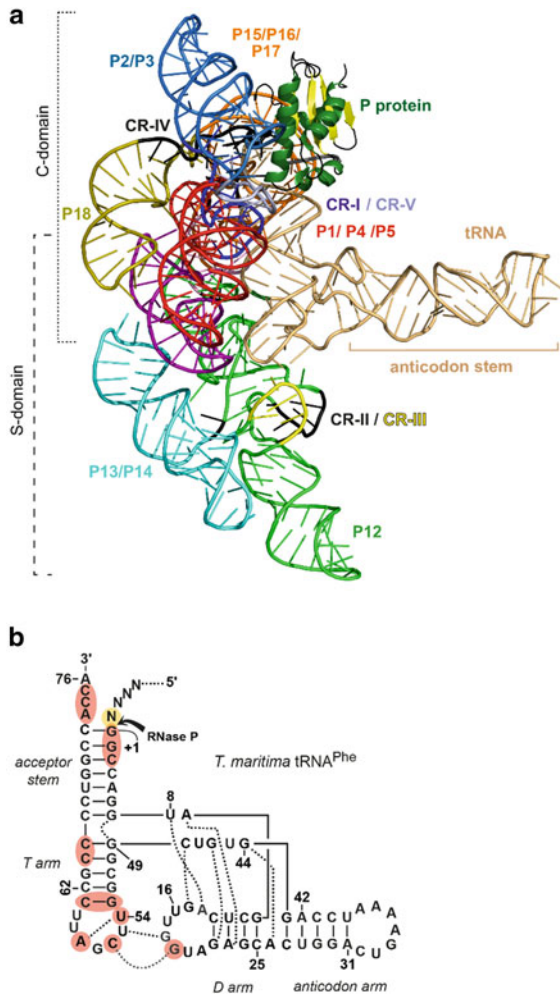


Fig. 11.3 (continued)



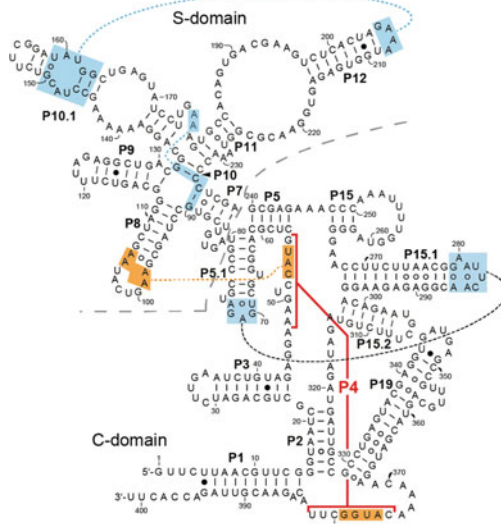
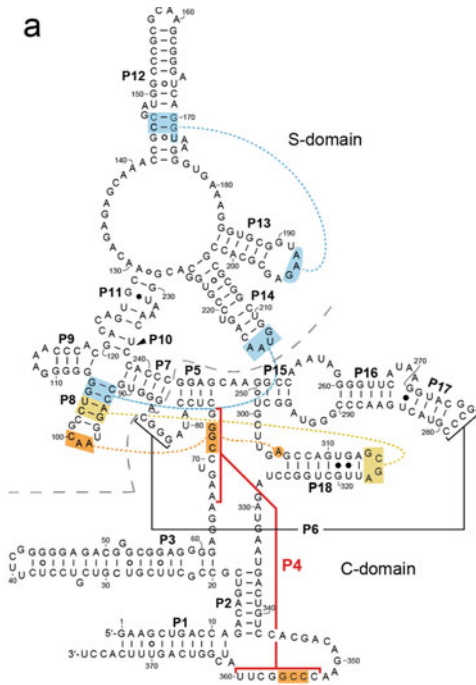
**Fig. 11.4** **a** Tertiary structure of the *T. maritima* RNase P holoenzyme in complex with tRNA (Reiter et al. 2010). The P RNA is colored as in Fig. 11.3b and brackets on the left delineate structural elements of the C-domain and S-domain. Helices P18 and P13/14 form a second layer that is piled onto the core helices. The P protein is shown in ribbon presentation, with  $\alpha$ -helices illustrated in dark green and  $\beta$ -strands in yellow. The *T. maritima* tRNA<sup>Phe</sup> with engineered anticodon arm extension (Reiter et al. 2010) is shown in sand color. The anticodon arm points away from the holoenzyme and only the coaxially stacked acceptor stem/T domain forms contacts with the P RNA (see panel b), primarily at the end of the acceptor stem and at the tRNA “corner” (where D and T arm interact) with T-loop structures formed by CR-II and CR-III of the P RNA. **b** Secondary structure of *T. maritima* tRNA<sup>Phe</sup>. Tertiary interactions, inferred from the crystal structure of yeast tRNA<sup>Phe</sup>, are indicated as dotted lines (adapted from Heide et al. 2001). Nucleotides making contacts to the P RNA are highlighted in pink. Nucleotide -1 is highlighted in yellow to indicate its involvement in metal ion coordination during catalysis

## 11.2.2 Archaeal RNase P: A Distinct and Increased Protein Moiety

Archaeal RNase P also contains a single P RNA subunit, but 5 different protein subunits (Fig. 11.2), which increases the protein content from ~10% in Bacteria to ~40%. The archaeal P RNA has been divided into three structural types: type A (ancestral), M (Methanococci), and T (Thermoproteaceae) (Fig. 11.5b). Type A is the most commonly found variant and resembles the bacterial type A, but lacks P18 and the region of P13/14 (Harris et al. 2001; Fig. 11.5). In bacterial P RNA, these elements, although neither involved in catalysis nor substrate interaction, engage in long-range tertiary interactions that stabilize the active fold of the RNA. Thus, archaeal type A RNAs have become more dependent on association with their protein subunits to adopt an enzymatically active conformation. This explains why archaeal type A RNAs are only weakly active *in vitro* in the absence of their protein cofactors, and this weak activity requires elevated salt concentrations such as 4 M  $\text{NH}_4^+$  and 0.3 M  $\text{Mg}^{2+}$  ions (Pannucci et al. 1999). It was further shown that substantial RNA-alone activity can be restored in an archaeal type A RNase P RNA by introducing minor changes toward the bacterial consensus into its C-domain (Li et al. 2009, 2011). The archaeal S-domain, however, has lost the capacity to support tight and productive substrate binding in the absence of protein cofactors (Li et al. 2009). A chimeric P RNA, consisting of the C-domain of *Methanothermobacter thermautotrophicus* P RNA (archaeal type A, with three minor alterations), the *E. coli* S-domain (bacterial type A) and further stabilized by implementing two interdomain contacts of bacterial P RNAs, was able to provide the essential RNase P function in *E. coli* cells (Li et al. 2011).

Euryarchaeota, such as the genera *Methanococcus* and *Archaeoglobus*, encode type M RNAs, which additionally lack P8 and everything distal to P15 including P16, P6 and L15 (Harris et al. 2001; Fig. 11.5). This removes the L8-P4 interdomain contact and abrogates base pairing of the tRNA 3'-NCCA end with the L15 loop, an important interaction in the bacterial system. These truncations increase the RNA's

**Fig. 11.5** Variation and conserved core structure of RNase P RNAs. Exemplaric secondary structures of RNase P RNAs. The central P4 helix is highlighted in red. **a** For the two bacterial P RNAs, additional details are shown: the gray dashed line demarcates the catalytic (C-) and specificity (S-) domains. Long-range tertiary interactions are depicted by boxes connected by dotted lines; intradomain contacts are highlighted in light blue, interdomain contacts in orange and ocher. For tertiary interactions in *E. coli* P RNA, see Brown et al. (1996), Massire et al. (1998), Marszalkowski et al. (2008b), and Reiter et al. (2010). Tertiary interactions for *B. subtilis* P RNA are illustrated according to the crystal structure of *G. stearotheermophilus* P RNA [L8:P4, L5.1:L15.1; (Kazantsev et al. 2005)] or to the *B. subtilis* S-domain structure (Krasilnikov et al. 2003). **b** Secondary structures of three archaeal and two eukaryal P RNAs. At the bottom on the right, the P4 consensus sequence/structure is shown, combining conserved features of bacterial and eukaryal nuclear RNase P RNAs (Marquez et al. 2005). Nucleotides highlighted in red are universally conserved, as are the highlighted base pairs which include G-A and A-C pairs in rare cases; base identities not highlighted are frequently but not always found at this position



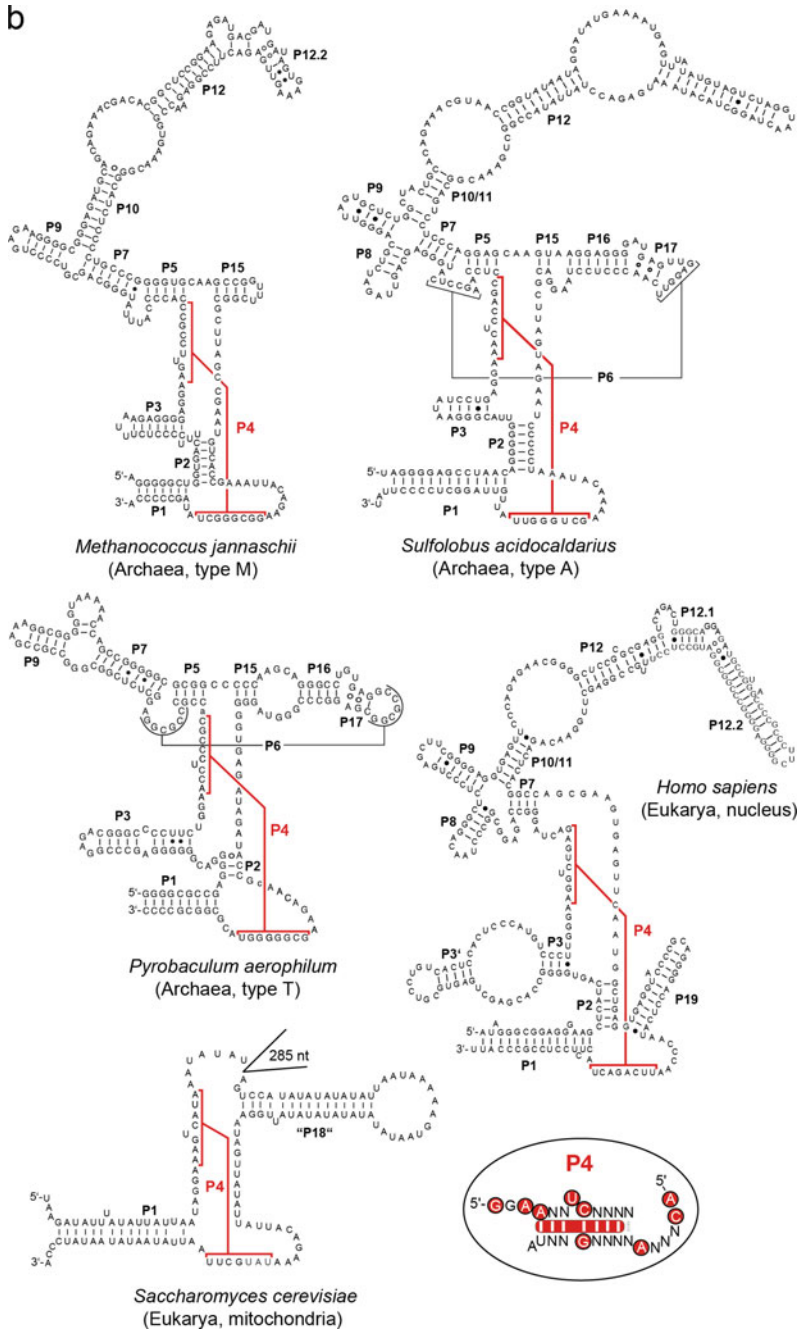


Fig. 11.5 (continued)

dependence on its protein cofactors for adopting an active fold. Accordingly, no RNA-alone activity has so far been observed for type M RNAs (Pannucci et al. 1999). P RNA of type T (Thermoproteaceae) was discovered in *Pyrobaculum* species and related crenarchaea; it is a minimized version of the RNA lacking almost the entire S-domain (Lai et al. 2010; Fig. 11.5). Yet, in contrast to type M, P15, P16, and P6 are present and may still allow base pairing with tRNA 3'-NCCA ends. This pygmy P RNA form displayed weak residual RNA-alone activity in vitro (Lai et al. 2010).

All five archaeal protein subunits, Rpp21, Rpp29, Rpp30, Pop5, and Rpp38/L7Ae, have homologs in eukaryal nuclear RNase P (Jarrous and Gopalan 2010), but none shows homology to the bacterial RnpA protein. The intensely studied Rpp21 of *Pyrococcus horikoshii* (~15 kDa) contains a C-terminal zinc ribbon domain that adopts an L-shaped overall structure with multiple basic amino acids on one face (Kakuta et al. 2005). Rpp29 (~11 kDa), which is a member of the oligosaccharide-binding fold family, has a structured  $\beta$ -barrel core and unstructured N- and C-terminal extensions (Boomershine et al. 2003; Sidote et al. 2004; Sidote and Hoffman 2003). X-ray analysis of *P. horikoshii* Rpp30 (~25 kDa) revealed that the protein consists of 10 helices (two of which have a surface enriched in positive charges) and seven  $\beta$ -strands, showing similarity to a TIM-barrel structure (Takagi et al. 2004). Pop5 (~10 kDa) folds into an  $\alpha$ - $\beta$  sandwich structure. Although not a homolog of the bacterial RnpA protein, some sequence similarities to the helix of RnpA containing the RNR-motif were observed (Wilson et al. 2006). L7Ae (~13 kDa), known to recognize K-turns, adopts an  $\alpha$ - $\beta$ - $\alpha$  sandwich fold (Suryadi et al. 2005).

Recently, a cryo-electron microscopy (cryo-EM) structure of the in vitro reconstituted RNase P holoenzyme of *Methanocaldococcus jannaschii* (type M) was solved at ~4 Å (Wan et al. 2019). Combined and supplemented with information from previous structural and biochemical studies, several architectural features have become evident. The entire holoenzyme adopts a dimeric conformation with two P RNAs, a (Pop5-Rpp30)<sub>2</sub> heterotetramer in the center and an Rpp29-Rpp21-Rpp38 heterotrimer at each end (Fig. 11.2b). Major contributions to our understanding of protein composition came from biochemical analyses of RNase P from Pyrococci (type A). First, NMR-binding studies revealed an interaction between Rpp21 and Rpp29 (Amero et al. 2008) and further details were revealed by an X-ray structure of this binary complex. N-terminal helices of Rpp21 interact with an N-terminal  $\beta$ -strand and a C-terminal helix of Rpp29 via hydrogen bonds and salt bridges. The complex harbors a positively charged cluster on one face probably involved in RNA interaction (Honda et al. 2008). Indeed, footprinting experiments showed binding of this binary complex to the S-domain of the P RNA (Xu et al. 2009). Functional reconstitution and NMR analyses led to the discovery of a second binary complex consisting of Pop5 and Rpp30 (Tsai et al. 2006; Wilson et al. 2006). X-ray analysis revealed a homodimer of *P. horikoshii* Pop5 in the center of a (Pop5-Rpp30)<sub>2</sub> heterotetramer, which led Makoto Kimura and coworkers to suggest the possibility of a dimeric (two P RNAs) holoenzyme structure (Kawano et al. 2006). Dimerization of Pop5 is mediated by hydrogen bonding interactions of the loops between the two N-terminal helices of each protein monomer. Each Pop5 molecule interacts with two Rpp30 molecules via hydrophobic interactions between the two N-terminal Pop5

helices and helix 7 in one and helix 8 in the other Rpp30 molecule. The surface of the heterotetramer has an unequal charge distribution with one face positively charged and the other either uncharged or negatively charged. Footprinting studies finally revealed binding of this complex to the C-domain of the P RNA (Tsai et al. 2006). Rpp21-Rpp29 binding to the S-domain increased substrate affinity, whereas the Pop5-Rpp30 complex enhanced catalytic efficiency (Pulukkunat and Gopalan 2008; Tsai et al. 2006). For the fifth archaeal RNase P protein, Rpp38/L7Ae, characterization using gel shift assays and chemical probing localized two possible P RNA-binding sites: the region of P15-17 in the catalytic domain and the P12 element which is part of the S-domain (Fukuhara et al. 2006).

In the recently published holoenzyme structure (Wan et al. 2019), the enzyme core is composed of the (Pop5-Rpp30)<sub>2</sub> heterotetramer and each Pop5 molecule binds to CR-IV (part of the active site) and P2-P3 of one P RNA component, whereas each Rpp30 monomer interacts with both P RNA molecules by contacting CR-V of one P RNA and P2-P3 of the other one (Fig. 11.2b). The binding region of the (Pop5-Rpp30)<sub>2</sub> complex is in close vicinity to the minimized P15 element in the type M RNase P cryo-EM structure, while it was inferred to contact the L15 region in type A RNA (Tsai et al. 2006). The Pop5-Rpp30 interaction with P RNA seems to contribute to stabilization of the active site and probably influences the correct position of catalytic Mg<sup>2+</sup> ions (Sinapah et al. 2011). Rpp30 contacts Rpp29 to bind the Rpp29-Rpp21-Rpp38 heterotrimer at each site of the (Pop5-Rpp30)<sub>2</sub> heterotetramer. This interaction, resulting in a protein decamer, has not been described previously. In line with previous footprinting assays, Rpp21 in particular forms strong interactions with the P RNA S-domain including the T-loops (Fig. 11.4a) and the P12 stem. As the archaeal RNA's S-domain is incapable of adopting a functional fold on its own to allow recognition of the D- and T-loop (one major interaction for substrate recognition in the bacterial enzyme), the Rpp21-Rpp29 complex is needed to structure and position the so-called T-loop anchor in the S-domain for interaction with the corner of the tRNA L-shape (Sinapah et al. 2011; Wan et al. 2019). In the cryo-EM structure (Wan et al. 2019), Rpp38/L7Ae recognizes the K-turn in P12.1.

In conclusion, the archaeal RNase P holoenzyme can form a dimer harboring two precursor tRNA-binding sites and two active sites. Yet, a detailed molecular understanding of substrate interaction and catalysis is not possible based on a structure of ~4 Å resolution and thus requires additional efforts. Furthermore, light needs to be shed on the architectural and functional differences between archaeal type A versus type M enzymes. Heterologous holoenzyme assemblies (type A RNA with type M proteins and vice versa) were active, but had lower activities than the corresponding homologous assemblies (Chen et al. 2010). This indeed points to differences between the two architectural enzyme types, attributable to co-evolution of type A RNA and cognate protein subunits on the one hand and type M RNA and cognate protein subunits on the other hand.



### 11.2.3 *Eukaryal Nuclear RNase P: Further Expansion of the Original Archaeal Protein Set*

The eukaryal RNA-based nuclear RNase P is composed of the conserved catalytic RNA subunit, homologs of the five archaeal proteins and four to five additional protein subunits. The protein content is thereby further increased to  $\sim 70\%$ . The RNA subunit, still having very weak but specific RNA-alone activity (Kikovska et al. 2007), and whose structural complexity is comparable to that of archaeal type M RNAs, has become even more protein-dependent than its archaeal counterparts. Phylogenetic studies on the RNA subunit revealed a conserved core including helical elements P1, P2, P3, P4, P7, P8, P9, P10, P11, and P12. CR I to V are identifiable as well in eukaryal P RNAs (Marquez et al. 2005). Some Eukarya-specific helices (eP8, eP9, eP15, eP19) are found in similar position to those in Bacteria, but have been designated “eukaryal” because of uncertain functional homology. Eukaryal P RNAs share only low overall sequence conservation. The differences, such as P12 extensions, are primarily in peripheral elements. Compared to the bacterial and archaeal RNAs, eukaryal P RNAs lack helix P5 and often any kind of P15 element, lack interdomain contacts, and have lost overall structural and conformational rigidity, especially in the P7–P10/11 region (Fig. 11.5). The P3 hairpin in bacterial and archaeal P RNAs is replaced by the helix–loop–helix subdomain P3, which is essential for protein binding (Esakova and Krasilnikov 2010; Frank et al. 2000; Marquez et al. 2005; Walker and Engelke 2006).

Concerning overall holoenzyme architecture, protein composition, and enzymatic activity and function, human and yeast enzymes have been studied most intensely. The protein moiety of human nuclear RNase P includes five proteins homologous to the archaeal enzyme (Pop5, Rpp29, Rpp30, Rpp21, and Rpp38/L7Ae) and five additional proteins (Pop1, Rpp25, Rpp20, Rpp14, and Rpp40) (Eder et al. 1997; Jarrous and Altman 2001; Jarrous 2002; Walker and Engelke 2006). The yeast enzyme harbors homologs to eight of these proteins (the archaeal core of five plus Pop1, as well as Rpp25 and Rpp20 homologs), but lacks Rpp14 and Rpp40 and instead includes the protein Pop8 (Chamberlain et al. 1998; Rosenblad et al. 2006; Walker and Engelke 2006). As most of the proteins of human and yeast nuclear RNase P were identified independently, several homologs carry different names (see Table 11.1 for comparison).

The proteins of the eukaryal RNP enzyme show a large variation in size (14–115 kDa), and they are in general quite basic ( $\text{pI} > 9$ ), with one exception each in the human and yeast enzyme (Rpp40 and Pop8). In 2012, a low-resolution (17 Å) electron microscopy structure of the yeast enzyme, using GFP-tagged protein subunits and a monoclonal GFP antibody to increase the size of particles, revealed some insights into the overall architecture (Hipp et al. 2012). Advanced structural insight was only gained more recently by cryo-EM structures of the yeast ( $\sim 3.5$  Å resolution; Lan et al. 2018) and the human RNase P holoenzyme ( $\sim 4$  Å resolution; Wu et al. 2018).

Overall, the human holoenzyme adopts an elongated conformation, with the RNA largely covered by protein, except for one subarea where the RNA is accessible

**Table 11.1** RNase P proteins of RNA-based enzymes across the domains of life. Proteins within the same row share sequence homology

Bacteria	Archaea	Eukarya	
		Yeast	Human
RnpA			
	Pop5	Pop5	Pop5
	Rpp29	Pop4	Rpp29
	Rpp30	Rpp1	Rpp30
	Rpp21	Rpr2	Rpp21
	Rpp38/L7Ae	Pop3	Rpp38
		Pop1	Pop1
		Pop6	Rpp25
		Pop7	Rpp20
			Rpp14
			Rpp40
		Pop8	

from one side (Wu et al. 2018; Fig. 11.2c). The protein moiety is composed of a single copy of the largest protein subunit (Pop1) and three subcomplexes: the Rpp20-Rpp25 heterodimer, Pop5-Rpp14-(Rpp30)<sub>2</sub>-Rpp40 heteropentamer, and an Rpp29-Rpp21-Rpp38 heterotrimer. The RNA subunit adopts a single layer conformation and interacts with all ten protein subunits (Fig. 11.2c). Pop1 forms extensive interactions with the C-domain, essentially embracing the C-domain from one side. Although the P3 element is in close vicinity, no direct contact of Pop1 and P3 was proposed, which is in contradiction to previous studies (Ziehler et al. 2001). Instead, the Rpp20-Rpp25 dimer seems to be responsible for interaction with this part of the P RNA, which is in line with more recent findings demonstrating that Pop6 and Pop7 interact with P3 of yeast P RNA (Esakova et al. 2008; Perederina et al. 2007, 2010). The heteropentamer leads to further stabilization of the C-domain, which is achieved by Rpp30, Rpp40, and Rpp14 contacting the region of P1, P2, and P19, while Pop5 and the second molecule of Rpp30 bind CR-IV and CR-V. Rpp29 seems to stabilize the connection of the C-domain and S-domain, as it interacts with P1 (C-domain) and P9 (S-domain). Rpp21 and Rpp38 contact and stabilize the S-domain by interacting with the P12 and CR-II/III region, where Rpp38 specifically recognizes the K-turn in P12.1.

The overall architecture of the yeast enzyme (Lan et al. 2018) resembles that of the human enzyme in the following aspects: the Pop5-Pop8-(Rpp1)<sub>2</sub> heterotetramer (corresponding to the Pop5-Rpp14-(Rpp30)<sub>2</sub>-Rpp40 heteropentamer in the human enzyme) contacts the C-domain including the catalytic core, Pop6-Pop7 binds the P3/P3' element, and Pop4-Rpr2-Pop3 interacts with both P RNA domains. Especially Pop4 seems to be a bridging element for stabilization of the connection between the C-domain and S-domain in the same way as Rpp29 in the human enzyme. Yeast

Pop1, which is shorter than its human counterpart, also forms extensive interactions with the C-domain, but its RNA-bound structure only partially overlaps with Pop1 in the human holoenzyme; this correlates with the architectural differences between the two P RNAs. In addition, yeast Pop1 helps to organize the yeast-specific multi-helix junction P7-P8'-P8-P9-P10/11 (Lan et al. 2018). In contrast to the human enzyme, Pop3 (corresponding to Rpp38 in human RNase P) makes only limited direct interactions with the P RNA.

The cryo-EM structure of human nuclear RNase P in complex with a tRNA (at 3.7 Å resolution; Fig. 11.2c) as well as the complex of the yeast nuclear enzyme and a pre-tRNA (3.5 Å) indicated stacking of the T-loop and D-loop with CR-II and CR-III, which is comparable to the bacterial enzyme (Wu et al. 2018; Lan et al. 2018). This is in line with the “measuring mechanism” proposed for cleavage-site selection by RNA-based eukaryal RNase P based on biochemical studies indicating that these enzymes have two major contact sites, the cleavage site and the corner of the tRNA L-shape, which have to be at an optimal distance to each other (Yuan and Altman 1995; Carrara et al. 1995). In the cryo-EM structures of the human and yeast nuclear RNase P, Pop1 was inferred to directly contact the acceptor stem of the tRNA substrate (Wu et al. 2018; Lan et al. 2018). This indicates that the protein moiety not only stabilizes the P RNA for catalysis, as in the case of archaeal RNase P, but also directly interacts with the substrate RNA, which can be considered as a further curtailing of the RNA's functional competence. Concerning 5'-leader binding, the protein subunits Pop5 and one Rpp1 in the yeast enzyme seem to be involved by forming a basic surface for electrostatic interaction (Lan et al. 2018).

#### ***11.2.4 On the Evolutionary Origin of P RNA and the Inflation of the Protein Moiety at the Expense of the RNA's Structural Integrity in Archaeal and Eukaryal RNA-Based RNase P***

RNA-based RNase P is generally assumed to be the ancient form of the enzyme. It is commonly described as one of the rare relics of a primordial “RNA world” where, like in its big siblings the spliceosome and the ribosome, the catalytic activity still resides in the RNA; in fact, among those modern RNA enzymes, bacterial P RNA is probably the catalytically most proficient one in the absence of its protein moiety. According to Maizels and Weiner (1994), P RNA could have evolved to release functional (catalytic) RNAs from the small 3'-terminal genomic replication tags (predecessors of tRNAs, before the advent of templated protein synthesis) required for their replication. Gray and Gopalan (2020) more recently developed an idea originally entertained by Altman and Kirsebom (1999) that such a trans-acting P RNA progenitor was probably essentially orthologous to the C-domain of modern P RNAs and might have in fact developed from a small cis-cleaving ribozyme initially attached to those replication tags. This idea is based on the observations that the

C-domain alone and even a 31-nt RNA derived from the C-domain are capable of catalyzing the canonical cleavage of pre-tRNA, albeit at lower efficiency (Green et al. 1996; Loria and Pan 1999; Tsai et al. 2006; Kikovska et al. 2012), and that P RNAs or their C-domains, when covalently tethered to pre-tRNAs or stem-loop model substrates, are able to release the latter at appreciable rates (Kikuchi et al. 1993; Frank et al. 1994; Kikuchi and Suzuki-Fujita 1995). The S-domain was accordingly suggested to have been acquired later, at the stage of LUCA only (Gray and Gopalan 2020).

The ancient status of the RNA enzyme is also consistent with the conserved structure of P RNAs, with its presence in the vast majority of bacteria and all known archaea, and with the proteins having apparently been recruited to the RNP independently in these two domains of life only after their split from LUCA (Hartmann and Hartmann 2003; Evans et al. 2006). Intriguingly, the bacterial P protein and the unrelated archaeal/eukaryal Pop5 nevertheless bind to roughly the same structural elements of their respective P RNAs, and, by interaction with the 5'-extension of the pre-tRNA, appear to also contribute in a possibly similar way to substrate recognition (Reiter et al. 2010; Lan et al. 2018; Wu et al. 2018; Wan et al. 2019). However, the detailed structural information on RNA-based RNase P from all three domains of life clearly confirmed that some of the long-range tertiary interactions of bacterial P RNAs that guarantee its stability required for substrate binding and catalysis are already replaced by proteins in archaeal RNase P. And this increase of the protein moiety at the expense of the RNA's structural integrity was apparently resumed in the early evolution of Eukarya, when the archaeal protein moiety was further expanded to become roughly doubled in mass in the derived RNA-based eukaryal nuclear RNase P enzymes, again accompanied by a further loss of stabilizing intramolecular RNA interactions. To explain this inflation of the protein moiety in the archaeal/eukaryal RNase P lineage, two opposing, but not mutually exclusive, hypotheses have been put forward, which are briefly reviewed in the following.

On the one hand, the increase in protein number and mass has been suggested to be the result of positive selection, with the recruitment of additional protein subunits explained by functional gains and/or increased versatility of the enzyme (Marvin and Engelke 2009; Jarrous and Gopalan 2010; Walker et al. 2010; Howard et al. 2013; Engelke and Fierke 2015). While no such additional substrates or functions have been reported for archaeal RNase P so far, nuclear RNase P in mammals and yeast has been reported to process non-tRNA substrates. In the case of mammals, these "extra" substrates are tRNA-like structures found in certain long non-coding RNAs (lncRNA) (MALAT1, MEN- $\beta$ ; reviewed in Jarrous and Gopalan 2010). While such tRNA mimics can obviously not explain the increased protein complexity as they would be substrates for basically any form of RNase P, the substrate spectrum for yeast nuclear RNase P was reported to go far beyond tRNA-like structures and to include numerous unspliced mRNAs, snoRNA precursors, and other non-coding RNAs (Coughlin et al. 2008; Marvin et al. 2011). However, using an RNase P-deficiency model different from the originally used temperature-sensitive *rpr1* (the gene encoding yeast nuclear P RNA) allele, we could not reproduce the accumulation of a selection of these non-tRNA candidate substrates (Weber et al. 2014). Moreover, yeast cells with the

genetic replacement of nuclear RNase P by *A. thaliana* PRORP3, a single-subunit protein-only form of RNase P, were viable and showed essentially unimpaired growth under a wide variety of conditions (Weber et al. 2014). Again, none of the proposed non-tRNA substrates of the RNP enzyme that we tested accumulated in the PRORP3-based strains, and their phenotype and fitness did not point to any functional deficit in RNA processing or any other vital function either. These results suggest that the spectrum of RNA substrates of nuclear RNase P in yeast is in fact much narrower than previously assumed and conceivably restricted to tRNAs and tRNA-like structures.

As an alternative hypothesis, we therefore suggested that the increasing number of proteins in the archaeal/eukaryal RNase P lineage may have co-evolved with the RNA's loss of structural integrity through a process called constructive neutral evolution (Weber et al. 2014). According to this concept (Stoltzfus 1999; Lukeš et al. 2011), a catalytically proficient P RNA "collected" RNA-binding proteins that fortuitously had the capacity to stabilize its active conformation, e.g., by simply reducing the electrostatic repulsion problem of the polyanionic RNA. This stabilization effect made the RNA more permissive toward the accumulation of structurally destabilizing mutations, which in turn increased the RNA's dependency on the presence of the initially facultative protein-binding partner(s) for adopting its catalytically active conformation. The cofactor dependency is assumed to increase in a ratchet-like process, where the progressive mutational destabilization of the RNA reinforces this mechanism while making the RNA's reversion to autonomy increasingly unlikely. This evolutionary model is also consistent with the recent cryo-EM structures of archaeal and eukaryal nuclear RNases P, which indicate that the comprehensive shell of protein cofactors primarily enables the RNA to adopt the conformation required for catalysis and specific recognition of tRNA substrates, a conformation that bacterial P RNA still achieves via stabilizing RNA-RNA interactions (Lan et al. 2018; Wu et al. 2018; Wan et al. 2019).

A neutral evolutionary acquisition and primarily structure-stabilizing role of the protein moiety is also consistent with the finding that even in the complex eukaryal RNP enzyme the substrate specificity (and thereby function) is still held by the RNA subunit (Kikovska et al. 2007). Moreover, the essentially same set of proteins is associated with the related RNP enzyme RNase MRP (Esakova and Krasilnikov 2010; Walker et al. 2010) and three of them (Pop1 and the Pop6-Pop7 heterodimer) are also found in the yeast telomerase RNP (Garcia et al. 2020). The endonuclease RNase MRP obviously originated from gene duplication of the ancestral RNase P RNA early in eukaryal evolution (as RNase MRP is found in all the supergroups), and obviously underwent neofunctionalization, whereby its substrate specificity changed, one of the archaeal P proteins (Rpp21) was lost and two novel small proteins were acquired, yet still eight of the proteins remained identical to the RNase P protein set (Esakova and Krasilnikov 2010; Walker et al. 2010). RNase MRP was shown to be involved in the processing of 5.8S rRNA and cell cycle-dependent cleavage of certain mRNAs, but its full functional spectrum is still unclear. In any case, RNase MRP does not process pre-tRNAs. If the protein subunits had a major role in conferring substrate specificity, the functions of the two RNP enzymes would be expected to either largely overlap, or the protein moieties would be expected to have diverged

more markedly during eukaryal evolution. As the opposite of the two scenarios appears to be the case, the functions of RNase P and MRP appear essentially defined by their RNA moieties. This does not exclude the direct involvement of some of the proteins to RNA binding, as, e.g., exemplified by Pop1 binding to the acceptor stem (Lan et al. 2018; Wu et al. 2018), but such sporadic protein contacts do not question the central role of the RNA subunits in determining substrate specificity. The exact roles of Pop1 and Pop6-Pop7 in yeast telomerase have not been defined so far (Garcia et al. 2020), yet it is conceivable that these proteins possess some general RNA structure-stabilizing function also employed in this RNP.

Studies of the Jarrous laboratory have linked human nuclear RNase P to chromatin structure and function (for review see Jarrous and Gopalan 2010; Jarrous 2017). However, based on the following reasoning and observations, it appears unlikely that such potential novel function of the nuclear RNP has contributed to increasing the protein moiety in the eukaryal lineage: While the protein moiety apparently evolved early in eukaryal evolution, a significant role of RNase P in chromatin remodeling and/or transcription appears not to be conserved as indicated by the loss of the RNA-based enzyme in several eukaryal lineages (see below) and by the lack of any substantial phenotype in yeast strains when the RNA-based enzyme was replaced by *A. thaliana* PRORP3 (Weber et al. 2014; see also above).

### ***11.2.5 Trends in the Evolution of Organellar RNA-Based RNase P***

Mitochondria are now firmly established to originate from the endosymbiotic acquisition of an ancestral  $\alpha$ -proteobacterium, a crucial event in eukaryal evolution that was associated with or took place close to their origin. The vast majority of present-day plastids are thought to have their single primary origin in the endosymbiosis of early eukaryal cells with a cyanobacterium at the root of modern Archaeplastida. In various other eukaryal phylogenetic groups, there is evidence for secondary endosymbiosis, where red and green algae were acquired as photosynthetic organelles, and also for serial secondary and tertiary endosymbiotic uptakes of algae in the dinoflagellates, which have an unusual diversity of plastids ascribed to their propensity to lose, substitute or gain new plastids (Keeling 2004). Following endosymbiosis, organellar genomes were generally destined to reduce their size and to lose many of their genes to the nuclear genome of the host. The bacterial-like RNase P protein subunit (RnpA) has been evolutionarily more volatile than the RNA subunit (RnpB): whereas an RnpA homolog has not been detected in any of the many sequenced organellar genomes so far, an RNA subunit conforming to the universally conserved P RNA core structure, though degenerated to almost unrecognizability in fungi/Ascomycota (see below), could be identified in organellar genomes. This includes the photosynthetic organelles of Glaucophyta (*Cyanophora paradoxa*), Rhodophyceae (red algae) and

some Chlorophyta (green algae) within the Archaeplastida, as well as mitochondrial genomes of fungi (patchy occurrence), Jakobida (Excavata) and some Chlorophyta (Mamiellophyceae, e.g., *Ostreococcus tauri*; see below) (see, for example, Seif et al. 2003; Lai et al. 2011; Lechner et al. 2015). Most, if not all, plastid-encoded P RNAs were identified in primary photosynthetic eukarya. No P RNA gene was found in mitochondrial genomes of Holozoa (supergroup Opisthokonta). The majority of mitochondrial P RNA genes were identified in fungal species (there usually termed *rpm1*), primarily in the phylum Ascomycota (Lechner et al. 2015).

The few protein subunit candidates of organellar RNA-based RNase P identified so far are encoded in the nucleus and are either descendants of bacterial-type RnpA proteins or belong to the group of Rpm2, which are pentatricopeptide repeat (PPR) proteins though unrelated to PRORP.

RnpA-like proteins were identified in the nuclear genomes of several Mamiellophyceae of the Chlorophyta subgroup of Archaeplastida (Lai et al. 2011; Lechner et al. 2015). These proteins were predicted to localize to mitochondria and chloroplasts and carry N- and C-terminal extensions that are absent from bacterial RnpA proteins (Lechner et al. 2015). Organellar RNase P has been studied to some extent in the Mamiellophyceae species *O. tauri* that encodes an active PRORP (Lai et al. 2011) predicted to localize to the nucleus (Lechner et al. 2015). In vitro, recombinant *O. tauri* RnpA was shown to form a functional holoenzyme with bacterial P RNA, but not with in vitro transcripts of the P RNAs encoded in the mitochondrial and chloroplast genomes of the same organism (Lai et al. 2011). A likely scenario is that these organellar P RNAs form complexes with the aforementioned enlarged RnpA proteins that are imported from the nucleus, but have recruited additional yet unidentified protein subunits to assemble functional RNase P holoenzymes. Of course, at present it cannot be excluded that nuclear encoded PRORPs of Mamiellophyceae might also be imported into organelles by utilization of alternative translation start codons. Nonetheless, recruitment of novel protein cofactors seems to be a common evolutionary trend in organellar RNase P evolution. Mitochondrial RNase P activity was reported to copurify with seven, as yet unidentified protein subunits in *Aspergillus nidulans* (Lee et al. 1996), reminiscent of the protein content of nuclear RNP RNase P. Another interesting example is the photosynthetic organelle (cyanelle) of the glaucophyte *C. paradoxa*. The cyanelle occupies a position intermediary between free-living cyanobacteria and plastids; the organellar genome codes for a bacterial A-type RNase P RNA that is an essential part of the organellar RNase P (Baum et al. 1996); the protein content was estimated to be ~80% (Cordier and Schön 1999). The cyanelle P RNA is A,U-rich and conformationally labile, but weak RNA-alone activity as well as formation of a functional holoenzyme with *E. coli* RnpA could be demonstrated (Li et al. 2007), thus testifying its bacterial origin. The cyanelle RNase P thus represented the first case of an organellar P RNA with documented ribozyme activity, which has recruited novel, yet unknown protein subunits for enzyme activity in vivo. The amoeba *Paulinella chromatophora* (SAR, Rhizaria) contains two blue-green photosynthetic, so-called chromatophores, an organelle derived from an evolutionarily more recent endosymbiosis of a cyanobacterium, thus representing a model of an earlier phase in plastid evolution. The gene content is ~25% of that of the most

closely related cyanobacterium. The cyanobacterial-like P RNAs of two *P. chromatophora* strains, which deviate from the bacterial consensus at only 2–3 positions, showed in vitro RNA-alone activity (although differing by a factor of ten) and formed a functional holoenzyme with a cyanobacterial RnpA protein (Bernal-Bayard et al. 2014). It will be interesting to find out if an *mnpA* gene is encoded in the nuclear genome and, if yes, whether the organellar P RNA forms a functional holoenzyme with such a yet to be identified RnpA protein and/or if new protein factors were recruited.

Bacterial-like RNase P RNAs, but no RnpA homologs, are encoded in mitochondrial genomes of jakobid flagellates (Excavata; Burger et al. 2013; Lechner et al. 2015). These P RNAs were reported to be catalytically inactive in vitro (Seif et al. 2006), suggesting they have also recruited novel protein cofactors that stabilize the RNAs' active fold. Jakobid mitochondrial genomes (~70–100 kb in size) impressively manifest the  $\alpha$ -proterobacterial descent of mitochondria. Beyond RNase P RNA, they code for a reduced form of tmRNA (translational control), a four-subunit bacteria-like RNA polymerase, which is replaced in other eukarya with a single-subunit phage-like enzyme encoded in the nucleus; moreover, protein-coding genes are preceded by potential Shine–Dalgarno translation initiation motifs (Burger et al. 2013).

A study based on activity assays and in silico structural comparison of plastid-encoded P RNAs from Rhodophyceae (red algae) and Chlorophyta (green algae), beyond confirming their lack of P RNA in vitro activity, attributed the inactivity to the absence of stabilizing intramolecular tertiary interactions, such as the L9-P1, L14-P8 and/or L18-P8 tetraloop-helix contacts, which support the RNA in adopting its functional conformation (Bernal-Bayard et al. 2014). This predicts the need for stabilization of plastid-encoded P RNAs by additional protein cofactors, a principle that was also operational in the evolution of archaeal and eukaryotic nuclear RNP RNase P.

Rpm2 (~119 kDa) was demonstrated to be an essential component of mitochondrial RNase P in *Saccharomyces cerevisiae* (Morales et al. 1992). Close Rpm2 homologs are restricted to the Saccharomycetales (Lechner et al. 2015). Rpm2 as part of native mitochondrial RNase P was reported to be organized in a stable complex in yeast mitochondria that comprises 136 proteins (Daoud et al. 2012). The complex contains seven proteins involved in RNA processing including RNase P and RNase Z, five out of six subunits of the mitochondrial RNA degradosome, and proteins involved in the fatty acid synthesis pathway, translation, metabolism, and protein folding. RNA components of the complex include the small and large subunit rRNAs and only ~70–90 nt long fragments of the yeast mitochondrial P RNA (Rpm1), consistent with the absence of intact Rpm1 RNA (427 nt) as reported earlier (Morales et al. 1989). The just described composition of the supercomplex also provides a framework that may give clues as to the intricate phenotypes of Rpm2 truncations, such as defects in mitochondrial protein synthesis and processing of Rpm1 transcripts without abrogating tRNA processing (Stribinskis et al. 2001a, b), or defective mitochondrial RNase P function upon disruption of any enzyme of the type II fatty acid synthesis pathway (Schonauer et al. 2008). Finally, a fraction of Rpm2 localizes to the nucleus and



activates transcription of components of the mitochondrial import apparatus and mitochondrial chaperones (Stribinskis et al. 2005). The authors proposed that Rpm2 is involved in mitochondrial biogenesis and critical for maintaining viability when cells lose their mitochondrial genome. Finally, Rpm2 was shown to be enriched in and stabilize cytoplasmic P bodies, suggesting a role in cytoplasmic mRNA storage and decay (Stribinskis and Ramos 2007).

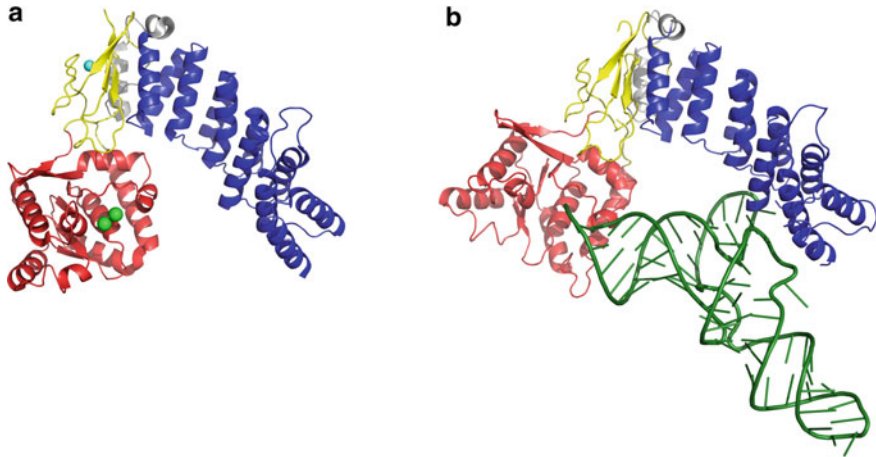
All aforementioned P RNAs encoded in organellar genomes are generally A,U-rich and thus conformationally labile, which explains their strong dependence on stabilization by protein cofactors to adopt a functional conformation, to coordinate catalytic  $Mg^{2+}$  ions and to specifically interact with tRNA substrates. P RNAs in fungi/Ascomycota (gene *rpm1*) are degenerated to the extreme. Identity elements are reduced to the terminal P1 helix and the catalytic core signatures CR-I, CR-IV, and CR-V (Seif et al. 2003; Fig. 11.5b). The RNA of *Kluyveromyces lactis* is already in the gray area, lacking a genuine P1 helix and deviating from the CR-I consensus, raising the possibility that some Ascomycota mitochondrial P RNAs may escape detection by the available search algorithms.

## 11.3 At Least Two Protein-Only Forms of RNase P Emerged Independently in Evolution

### 11.3.1 PRORP: A Protein-Only RNase P Invented in the Early Evolution of Eukarya

PRORP, an RNase P in most instances composed of a single polypeptide of about 60-kDa only, is found in the nuclei, mitochondria, and/or chloroplasts of many eukarya (Holzmann et al. 2008; Gobert et al. 2010; Gutmann et al. 2012; Taschner et al. 2012; Lechner et al. 2015; Bonnard et al. 2016). The protein is characterized by a three-domain architecture, where a pentatricopeptide repeat (PPR) domain is fused to a characteristic PIN-like domain (PiIT N-terminal) via a split zinc-binding domain (Fig. 11.6a). While PIN-like domains are found in all domains of life, the PRORP-specific nuclease fold and PPR domains are exclusively encoded in eukaryal genomes, but not in bacterial or archaeal ones. In the animal mitochondrial phylogenetic branch, PRORP requires two additional proteins for efficient catalysis (Holzmann et al. 2008), and the specific structure, function, and evolution of this multi-subunit form of eukaryal protein-only RNase P shall thus be reviewed in a separate section below.

The C-terminal catalytic domain of PRORPs was until recently assigned to the NYN family within the PIN domain-like superfamily of nucleases (Anantharaman and Aravind 2006). However, the catalytic domain of PRORPs is characterized by a specific arrangement of conserved aspartates and a histidine in two sequence signatures (parts 1 and 2; Lechner et al. 2015) that allows distinguishing them from other NYN-domain proteins, and based on the structure of the catalytic domain, PRORPs



**Fig. 11.6** **a** Crystal structure of *A. thaliana* PRORP1 (Howard et al. 2012; PDB: 4G24). The catalytic NYN domain is shown in red, the Zinc-binding domain in yellow and the PPR domain in dark blue; the zinc ion is depicted as a cyan sphere and the two putative catalytic metal ions ( $Mn^{2+}$ ) as green spheres. **b** Model of *A. thaliana* PRORP2 in complex with tRNA (green) in solution based on biochemical data and SAXS constraints (Pinker et al. 2017), kindly provided by Philippe Giegé, Strasbourg, France. The PPR domain interaction with the conserved elbow region of tRNAs approximates the recent crystal structure of a truncated PRORP1 PPR domain in complex with yeast tRNA<sup>Phe</sup> (Teramoto et al. 2020)

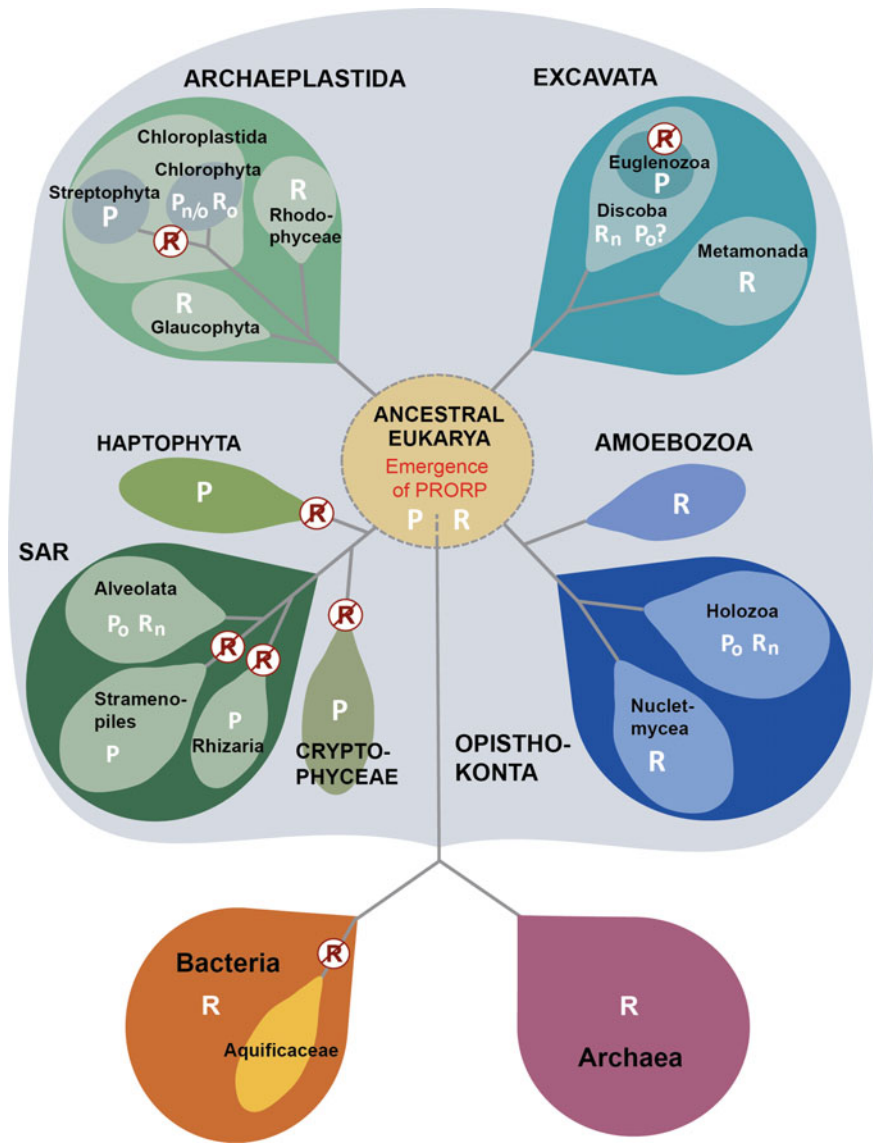
were recently suggested to comprise a separate group within the PIN domain-like superfamily distinct from the NYN group (Matelska et al. 2017; Gobert et al. 2019). Structurally, the catalytic domain of PRORP adopts the Rossmannoid fold of a central  $\beta$ -sheet sandwiched by  $\alpha$ -helices at both sides that is characteristic of all PIN-like domains (Howard et al. 2012; Matelska et al. 2017). Four aspartates, strictly conserved in the NYN domains of all PRORPs, are involved in the coordination of two catalytic metal ions, presumably magnesium (Howard et al. 2012).

The N-terminal  $\alpha$ -superhelical domain typically comprises a tandem array of five PPR or PPR-like motifs (Howard et al. 2012). While highly degenerate in sequence, the characteristic array of helix-turn-helix motifs of PPRs is generally recognizable (Lechner et al. 2015). PPR proteins are widely distributed in Eukarya and massively expanded in land plants (Small and Peeters 2000; Lurin et al. 2004; Schmitz-Linneweber and Small 2008). They are generally involved in RNA metabolism and the 35-amino acid motif is considered to bind to a specific nucleobase in a target RNA according to a “code” specified by amino acid positions 5 and 35 of a given PPR repeat (also referred to as 4/34 and 6/1’, depending on the numbering system used) (Barkan et al. 2012; Yagi et al. 2013). However, the simple rules of this code do not apply to the PPR motifs in PRORPs (Brillante et al. 2016), but their PPR domain binds to the D/T-loop region of pre-tRNAs via different mechanisms, as recently revealed by the co-crystal structure of tRNA in complex with a PRORP PPR domain (Teramoto et al. 2020).

In the overall  $\Delta$ -shaped structure of PRORP, the N- and C-terminal domains form the two legs of the  $\Delta$  connected via a split zinc-binding domain at the apex (Howard et al. 2012; Gobert et al. 2013; Fig. 11.6). The first element of the zinc-binding domain is inserted in between the PPR and NYN domain and comprises a highly conserved CxxC motif; the two cysteines take part in coordinating the zinc ion (Fig. 11.6a). Other elements are found downstream of the NYN domain close to the C-terminus and consists of a (W/Y/F)HxPx and a (W/F)xCx2-3(R/K) sequence motif, with histidine and cysteine providing the remaining side chains for zinc coordination. The central structural zinc ion appears to be important for the proper orientation of the nuclease domain relative to the PPR domain. Pre-tRNA substrates are believed to dock to the inside of the  $\Delta$ , with the PPR domain facing the angle of the L-shaped tRNA and the nuclease domain the cleavage site Fig. 11.6b; Pinker et al. 2017), in line with the aforementioned PPR domain-tRNA co-crystal structure (Teramoto et al. 2020). Finally, optional organellar targeting signals and/or nuclear localization signals are found at the N-terminus of many PRORPs (Lechner et al. 2015).

While PRORPs are only found in Eukarya and exclusively encoded in nuclear genomes, a thorough phylogenetic analysis of available genomes showed that PRORPs are widespread in this domain of life and found in four of the five supergroups (Lechner et al. 2015; Fig. 11.7); no PRORP was identified in Amoebozoa. A search for the presence of different forms of RNase P and their sequence analysis moreover revealed that PRORPs can apparently be localized to any genetic subsystem (nucleus or organelle) of eukaryal organisms/phyla (Lechner et al. 2015). While PRORPs have apparently conquered all tRNA-processing compartments and completely replaced RNP forms of RNase P in land plants, stramenopiles, or trypanosomes, they are found in either only the nucleus or only the organelle of other eukarya, with a RNP RNase P found in the other compartment, respectively. And although even multiple copies of PRORP occasionally appear to coexist within a genetic subsystem (nucleus or organelle) of certain organisms/phyla, no evidence was found for the coexistence of PRORP and RNP RNase P within the same compartment, i.e., the occurrence of the two conceptually different forms of RNase P appears mutually exclusive in present-day Eukarya. Major phyla without PRORP proteins, apart from Amoebozoa, are fungi and basal primary photosynthetic groups such as Glaucophyta and Rhodophyceae.

A phylogenetic analysis of ~90 PRORP sequences representative of all major eukaryal groups using the maximum likelihood method (Lechner et al. 2015) suggested an early origin of the protein, apparently before eukaryal radiation. There is no evidence that its “invention,” as defined by the fusion of its characteristic PPR, bipartite zinc-binding and nuclease domains, occurred more than once in evolution (Lechner et al. 2015). The evolutionary origin of PRORP at the root of eukaryal evolution is supported by the protein’s wide distribution among Eukarya combined with the clustering of PRORP sequences in clearly distinct groups for Opisthokonta, Archaeplastida, Excavata, and for a major group of stramenopiles and Rhizaria species (Lechner et al. 2015). However, the analysis also suggests some horizontal gene transfer (e.g., by secondary or tertiary endosymbiosis) in stramenopiles (SAR), where species frequently have more than one PRORP, which cluster in evolutionary



**Fig. 11.7** Distribution of RNA-based and protein-only RNase P enzymes in the three domains of life (adapted from Lechner et al. 2015, and extended to Archaea and Bacteria). R and P designate RNA-based and protein-only RNases P, respectively; subscripts: n, nucleus; o, organelle; n/o, found in the nucleus and organelle(s). Crossed out R marks putative evolutionary events that resulted in the loss of (nuclear) RNP RNase P. The question mark indicates an example where limited genomic data prevented conclusions as to the occurrence of the given enzyme type in the respective group (Lechner et al. 2015). The diagram highlights how the distribution of RNase P seemingly involved multiple events of losses of either PRORP or RNP RNase P

distinct groups (Lechner et al. 2015). An example are Peronosporomycetes (SAR, stramenopiles), “fungal-like” organisms that lost the plastid acquired by secondary endosymbiosis, where one type of PRORP is predicted to be targeted to the nucleus and the second, separately clustering one, to mitochondria.

Subsequent to eukaryal radiation, PRORP was apparently either lost, or took over nuclear and/or organellar RNase P function from the ribonucleoproteins, leading in turn to the subsequent loss of the latter (Fig. 11.7). The relative ease of such an evolutionary swap was demonstrated by the experimental replacement of the endogenous RNP enzymes of *E. coli* or *S. cerevisiae* by various PRORP isoforms (Göbringer et al. 2017; Weber et al. 2014). Notably, in the eukaryal context of *S. cerevisiae*, such experimental swap was apparently possible without any obvious costs (Weber et al. 2014; see above for additional details). Moreover, in terms of the evolutionary plasticity of subcellular localization or horizontal gene transfer, a single-subunit protein enzyme is obviously more flexible than a multi-subunit enzyme, particularly a RNP. Thus, while PRORPs apparently invaded all subcellular tRNA-processing compartments, there is no evidence that any archaeal-type nuclear or endosymbiont-derived bacterial-type organellar RNP RNase P ever made it to another subcellular compartment during eukaryal evolution (Lechner et al. 2015). In contrast, e.g., a single PRORP gene encodes all three subcellular RNase P forms of *Chlamydomonas reinhardtii* (Bonnard et al. 2016). Consistently with the above said, PRORP also appears to have been (re-)introduced into some lineages by the horizontal gene transfer associated with secondary and tertiary endosymbiosis, e.g., in stramenopiles, as outlined above.

Finally, another aspect of PRORP evolution is worth mentioning: two PRORPs are expressed in *T. brucei* (Excavata), one localizes to the nucleus and the second to mitochondria (Taschner et al. 2012). However, the mitochondrial genomes of trypanosomatids lack any genes for tRNAs, which are imported in their mature form from the cytosol, raising the question what the function of the mitochondrial PRORP enzyme might be. Curiously enough, this mitochondrial PRORP nevertheless has RNase P activity in vitro (Taschner et al. 2012).

### ***11.3.2 Recruitment of Further Subunits to PRORP in Animal Mitochondria***

PRORP, first identified as the catalytic subunit of human mitochondrial RNase P, was originally termed mitochondrial RNase Protein 3 (MRPP3), because it was the last component of the tripartite enzyme to be identified (Holzmann et al. 2008). The other two components are TRMT10C (tRNA methyltransferase 10C; originally MRPP1) and SDR5C1 (short chain dehydrogenase/reductase family 5C, member 1; originally MRPP2). All three proteins are required for the reconstitution of robust RNase P activity in vitro (Holzmann et al. 2008; Vilaro et al. 2012), and the knockdown of any of them results in the accumulation of mitochondrial tRNA precursors (Holzmann

et al. 2008; Lopez Sanchez et al. 2011; Deutschmann et al. 2014; Rackham et al. 2016). However, rather than being merely a three-subunit proteinaceous RNase P, human mitochondrial RNase P turned out to be a multi-enzyme assembly, in which the nuclease PRORP became dependent on a tRNA methyltransferase complex rather unusual in itself (Holzmann et al. 2008; Vilardo et al. 2012). Although this form of RNase P was mainly studied in human cells, its components appear to be conserved throughout the metazoan lineage, and a functional characterization in *Drosophila* confirmed the essential role of all three subunits for mitochondrial RNase P function (Sen et al. 2016).

The involvement of SDR5C1 in mitochondrial RNase P function is still most puzzling. Being a member of the short-chain dehydrogenase/reductase (SDR) superfamily, a large group of nicotinamide adenine dinucleotide (phosphate)-dependent oxidoreductases (Kallberg et al. 2002), it serves as the mitochondrial short-chain l-3-hydroxy-2-methylacyl-CoA dehydrogenase that catalyzes the penultimate step in the  $\beta$ -oxidation of branched- and short-chain fatty acids and isoleucine (Luo et al. 1995). According to its relationship to 17- $\beta$ -hydroxysteroid dehydrogenases, its gene was named *HSD17B10* (Adamski and Jakob 2001). It was furthermore reported to be active on a wide range of alcohols and hydroxysteroids in vitro (He et al. 1999, 2000a, b, 2005a, b; Shafiqat et al. 2003), was identified as an amyloid- $\beta$ -binding protein (Yan et al. 1997, 2000; Lustbader et al. 2004), and can consequently be found under a plethora of names in the scientific literature, yet the biological relevance of most of these findings requires further scrutiny. Mutations in its gene cause a mitochondrial disease originally called 2-methyl-3-hydroxybutyryl-CoA dehydrogenase (MHBD) deficiency (Ofman et al. 2003; Zschocke et al. 2000)—now termed HSD10 disease—that is characterized by progressive neurodegeneration and cardiomyopathy (for review see Zschocke 2012). Paradoxically, it appears to be the protein's function in tRNA biogenesis rather than its dehydrogenase activity that is essential and, if defective, the cause of mitochondrial disease (Rauschenberger et al. 2010; Deutschmann et al. 2014; Vilardo and Rossmannith 2015; Chatfield et al. 2015). We therefore advocate the use of the systematic name SDR5C1 for this multifunctional protein, as it does not refer to any specific of its diverse activities or functions (Persson et al. 2009).

SDR5C1 is a polypeptide of 27 kDa forming a homotetramer of known crystal structure (Powell et al. 2000; Kissinger et al. 2004). An (undetermined) fraction of mitochondrial SDR5C1 forms a stable complex of 4:2 stoichiometry with TRMT10C (Holzmann et al. 2008; Vilardo et al. 2012; Vilardo and Rossmannith 2015; Oerum et al. 2018), presumably with two TRMT10C monomers symmetrically attaching to the SDR5C1 tetramer. This complex constitutes the methyltransferase responsible for  $N^1$ -methylation of purines at position 9 ( $m^1R9$ ) of mitochondrial tRNAs (Vilardo et al. 2012). The role of SDR5C1 within this complex, however, appears merely that of a scaffold, and neither its dehydrogenase activity nor an intact  $NAD^+$ / $NADH$ -cofactor-binding site are required to “support” the methyltransferase activity of TRMT10C or the endonucleolytic activity of PRORP (Vilardo et al. 2012; Vilardo and Rossmannith 2015).

TRMT10C belongs to the TRM10 group within the SpoU-TrmD (SPOUT) superfamily, a class of *S*-adenosyl methionine-dependent methyltransferases defined by a deep trefoil knotted catalytic domain (Anantharaman et al. 2002; Tkaczuk et al. 2007; Krishnamohan and Jackman 2019). The generally monomeric TRM10 enzymes (most other SPOUT methyltransferases are dimers) are responsible for the  $N^1$ -methylation of guanosine and/or adenosine at position 9 ( $m^1G9$ ,  $m^1A9$ ) of tRNAs in Archaea and Eukarya. However, while  $m^1G9$  is widely found in cytosolic tRNAs of Eukarya,  $m^1A9$  appears rare among those, and, with the exception of animal mitochondria, the two modifications have not been found in organellar tRNAs (consistent with their absence in Bacteria) (Jühling et al. 2009). The tRNA  $m^1R9$  modification is in fact nearly ubiquitous in animal mitochondria, i.e., in all 19 of the 22 human mitochondrial tRNAs that contain a purine at position 9, this purine is methylated to  $m^1A9$  or  $m^1G9$ , respectively (Suzuki and Suzuki 2014; Clark et al. 2016; Vilardo et al. 2020). Consistently, the TRM10 family has been expanded to two or three isoenzymes in the metazoan lineage (Jackman et al. 2003), with one of the homologs (called TRMT10C in vertebrates) apparently targeted to mitochondria (Holzmann et al. 2008). In contrast to the cytosolic and archaeal members of the TRM10 family, which are active as monomers on their own (Jackman et al. 2003; Kempnaers et al. 2010; Vilardo et al. 2012, 2020; Swinehart et al. 2013; Howell et al. 2019), TRMT10C strictly requires the above-mentioned interaction with SDR5C1 to show appreciable methyltransferase activity (Vilardo et al. 2012; Vilardo and Rossmanith 2015). However, as mentioned before, SDR5C1 appears to neither contribute to substrate binding nor catalysis (Vilardo et al. 2012), and the methyltransferase domain of TRMT10C resembles that of monomeric TRM10 enzymes (Oerum et al. 2018), thus providing no clue to the essential cofactor function of SDR5C1. Recent findings suggest that SDR5C1 stabilizes the binding of TRMT10C to pre-tRNAs (unpublished results), but the extent of this effect does not appear sufficient to explain the crucial role of SDR5C1 for the methyltransferase activity of TRMT10C.

The contribution of the TRMT10C-SDR5C1 methyltransferase complex to the RNase P activity of human PRORP, likewise, is also largely unclear. While the complex is definitely able to bind pre-tRNAs, its methyltransferase activity is not required to support pre-tRNA cleavage by PRORP, and methylation and cleavage also appear to be independent and, at least in vitro, to proceed uncoupled (Vilardo et al. 2012). Moreover, the TRMT10C-SDR5C1 complex is also involved in the cleavage of mitochondrial pre-tRNAs that are not even methylated at position 9, like tRNA<sup>Met</sup> and tRNA<sup>Ser(UCN)</sup> (unpublished results). Nevertheless, the methyl group donor *S*-adenosyl methionine was reported to stimulate cleavage in the case of some pre-tRNAs (independently of actual substrate methylation) (Karasik et al. 2019). Strangely enough, the TRMT10C-SDR5C1 complex was also reported to act as a general mitochondrial tRNA-processing platform, stimulating the cleavage activity of RNase Z (ELAC2) similarly like that of PRORP (Reinhard et al. 2017); however, we could not reproduce this effect and rather observed an inhibitory effect of the TRMT10C-SDR5C1 complex on RNase Z activity (unpublished observation).

In terms of structure, human PRORP does not appear to revealingly differ from single-subunit PRORPs, either. The crystal structures of two N-terminally truncated

human PRORP constructs showed a  $\Lambda$ -shaped structure comparable to *Arabidopsis* PRORP1 and PRORP2, but with substantial disorders in the NYN domain (Reinhard et al. 2015; Li et al. 2015). Both groups interpreted this as an inactive conformation unable to bind the catalytic metal ions and suggested that the TRMT10C-SDR5C1 complex would act in remodeling the NYN domain to enable the coordination of  $Mg^{2+}$  ions in the active site. However, both crystalized PRORP fragments were catalytically inactive even in the presence of TRMT10C-SDR5C1. Moreover, we recently found that human PRORP alone is in fact able to bind pre-tRNAs, to properly coordinate  $Mg^{2+}$  ions in its active site, and to even cleave some pre-tRNAs with considerable yet reduced efficiency, all in the absence of the other subunits (unpublished results). The TRMT10C-SDR5C1 complex appears to primarily, and in most cases dramatically, increase the cleavage rate of PRORP though.

So why does human PRORP require the TRMT10C-SDR5C1 methyltransferase complex if not for its methylating activity (see above), whereas (all) other PRORPs act on their own? And along the same lines, why does TRMT10C depend on the “scaffolding” SDR5C1, whereas (all) other methyltransferases of the TRM10 family do not require any partner protein? Obviously, in both cases the monomeric forms are ancestral, and the specific interactions and dependencies evolved during the (early) evolution of metazoan mitochondria. Now remarkably, the tRNAs of metazoan mitochondria are peculiar too. While the mitochondrial tRNAs of most eukarya conform to the ubiquitous, canonical type of tRNA structure, the mitochondrial tRNAs of metazoa notably deviate from this consensus structure. They are generally smaller, have often highly reduced, or even lack, D or/and T domains, and are frequently not able to form the typical tertiary interactions that stabilize the tRNA L-shape (Giegé et al. 2012; Suzuki et al. 2011; Watanabe et al. 2014; Wende et al. 2014). Moreover, a given set of metazoan mitochondrial tRNAs can be of pretty divergent structure, ranging from fully canonical to highly reduced. These peculiarities apparently also required some co-evolution of the tRNA-interacting factors (Watanabe et al. 2014; Jühling et al. 2018), and such co-evolution has been hypothesized to underlie the unique evolution of metazoan mitochondrial RNase P (Holzmann et al. 2008). Indeed, various other RNase P enzymes, including single-subunit PRORPs, were found to cleave at least some human mitochondrial tRNAs in vitro either inefficiently and/or aberrantly, or not at all (Rossmann et al. 1995; Karasik et al. 2019). The promiscuous methyltransferase (acting on all mitochondrial tRNAs with a purine at position 9), which was apparently acquired in early metazoan evolution as well (m1G9/m1A9 or TRM10-type enzymes have thus far not been found in mitochondria of other eukarya), could have evolved to become a general accessory factor for PRORP that safeguards proper cleavage of the structurally degenerated pre-tRNA substrates, a quasi-extra recognition handle beyond the direct pre-tRNA contacts of PRORP. Such cooperation might have loosened the structural constraints for PRORP binding to broaden the enzyme’s pre-tRNA substrate spectrum without putting unwanted RNAs at risk of being cleaved by a nuclease of loosened specificity. In fact, rather than merely helping with substrate binding, the TRMT10C-SDR5C1 complex seems to stimulate cleavage through conformational changes in the pre-tRNA substrate that facilitate phosphodiester hydrolysis by PRORP (unpublished results). The role of the



TRMT10C-SDR5C1 complex nevertheless also appears to involve specific protein-to-protein interactions with PRORP, as it cannot be replaced as RNase P cofactor by related TRM10 enzymes, although they are well capable of methylating mitochondrial tRNAs *in vitro* and can thus be assumed to rearrange the tRNA structure in a similar way (unpublished results); likewise, TRMT10C-SDR5C1 does not stimulate the cleavage activity of single-subunit PRORPs (unpublished results).

Whether or how the evolution of the TRMT10C-SDR5C1 complex can be related to the non-canonical and variable mitochondrial tRNA structures, appears less clear. As mentioned, monomeric TRM10 enzymes do methylate mitochondrial tRNAs *in vitro*; although only the formation of m<sup>1</sup>G9 was tested, the ability to also form m<sup>1</sup>A9 is unlikely associated with SDR5C1, as such dual, relaxed specificity is also found in some (monomeric) archaeal members of the TRM10 family, which anyway contains G9-, A9-, and R9-specific members (Jackman et al. 2003; Kempnaers et al. 2010; Vilardo et al. 2012, 2020; Singh et al. 2018; Howell et al. 2019). A possible evolutionary scenario appears to be the constructive neutral acquisition of SDR5C1 by TRMT10C. According to this concept of evolution (Stoltzfus 1999; Lukeš et al. 2011), a catalytically proficient predecessor of TRMT10C fortuitously “picked” SDR5C1 as a neutral binding partner, i.e., without functional consequences for any of them. This interaction stabilized the structure of the TRMT10C predecessor and allowed the accumulation of destabilizing mutations, whereby its methyltransferase activity became (increasingly) dependent on the interaction. Thereby the process started to progress in a ratchet-like manner, as the accumulation of further destabilizing mutations became increasingly more likely than reversion to autonomy, and SDR5C1 finally ended up as a structure-stabilizing scaffold of TRMT10C. SDR5C1 itself probably only escaped this ratchet because a major fraction of SDR5C1 acted unbound by TRMT10C all the time. This view is consistent with the observation that the enzymatic activity of SDR5C1 appears (still) independent of the interaction with TRMT10C, i.e., it is neither inhibited nor stimulated by the latter (Oerum et al. 2018).

In summary, neutral and coevolutionary adaptive processes appear to have shaped the peculiar form of proteinaceous RNase P in animal mitochondria.

### ***11.3.3 A Minimal Protein-Only RNase P in Bacteria and Archaea***

The bacterial phylum Aquificae includes the orders Aquificales (families Aquificaceae and Hydrogenothermaceae) and Desulfurobacteriales (family Desulfurobacteriaceae) (Gupta and Lali 2013). The Aquificae comprises thermophilic to hyperthermophilic gram-negative, motile and non-sporulating bacteria that inhabit terrestrial and marine hot springs worldwide (see Wäber and Hartmann 2019, and references therein). The Aquificaceae members *Aquifex aeolicus* and *Aquifex pyrophilus* grow

at temperatures of up to 95 °C, thereby belonging to the most thermophilic bacteria known to date.

The genome sequence of *A. aeolicus* (1.55 Mbp) published in the late nineties (Deckert et al. 1998) proposed a conundrum to the RNase P research community: neither an *rnpA* gene for the RNase P protein subunit nor the *rnpB* RNA gene could be identified in the densely coding *A. aeolicus* genome, despite numerous bioinformatic efforts worldwide. The surprise effect becomes evident when considering the conserved synteny of the bacterial *rnpA* gene: it is located close to and in most cases even cotranscribed with the neighboring *rpmH* gene coding for ribosomal protein L34 (Hartmann and Hartmann 2003), a constellation even found in *Hydrogenothermaceae* family members that are close relatives of *Aquifex* (Marszalkowski et al. 2008a). However, no *rnpA* and only the *rpmH* gene is present in *A. aeolicus* and *A. pyrophilus* at this location, although the overall synteny beyond *rnpA* in this genome region is conserved between the two hyperthermophilic *Aquifex* species and other more distantly related bacteria (Hartmann and Hartmann 2003; Marszalkowski et al. 2006, 2008a). Improved search algorithms for P RNA genes likewise failed to extract an *rnpB* gene candidate from the *A. aeolicus* genome (Li and Altman 2004). Nonetheless, the genetic organization of tRNAs (in tandem clusters and as part of ribosomal operons) on the one hand, and the presence of tRNAs with canonical mature 5'-ends in total RNA extracts from *A. aeolicus* on the other hand, implied the existence of a tRNA 5'-maturation activity, although RNase P activity could not be demonstrated for *A. aeolicus* cell extracts in initial attempts (Willkomm et al. 2002). The issue gained momentum when we (Marszalkowski et al. 2008a, b) and another laboratory (Lombo and Kaberdin 2008) contemporaneously detected RNase P activity in cell extracts of *A. aeolicus*. However, the identity and biochemical composition of RNase P in *A. aeolicus* remained enigmatic until its disclosure in 2017 (Nickel et al. 2017). In the latter study, RNase P activity could finally be assigned to a single polypeptide of ~23 kDa encoded by the gene Aq\_880. The small protein has several active site aspartates involved in the coordination of one or two catalytic metal ions and bioinformatic analysis revealed a position within the PIN domain-like superfamily; however, Aq\_880 is assigned to a different subgroup (PIN\_5 cluster, VapC structural group) than PRORP (Matelska et al. 2017). The recombinant protein catalyzes RNase P-specific tRNA 5'-end maturation with kinetics in the range of other RNase P enzymes. In gel filtration experiments, Aq\_880 elutes as a ~420-kDa complex, consistent with large homo-oligomeric complexes (three hexamers or six trimers). Some recombinant Aq\_880 even appears as a ~70 kDa band in SDS gels, confirming the protein's propensity to form stable oligomers (Nickel et al. 2017); oligomerization is a prominent mechanism of protein thermostabilization (Wäber and Hartmann 2019, and refs. therein). Finally, Aq\_880 was able to rescue the growth of an *E. coli* strain under conditions of a lethal knockdown of its endogenous RNase P. Even more surprisingly, a yeast strain with inactivation of its complex nuclear RNP RNase P could be rescued as well (Nickel et al. 2017). In both cases, growth efficiency was rescued to levels below that of the corresponding wild-type strain, indicating that Aq\_880 functions suboptimally in the heterologous hosts.

Bioinformatic searches then sporadically identified homologs of Aq\_880, termed HARPs for Homologs of Aquifex RNase P, in 5 other of the 36 bacterial phyla beyond Aquificae: a few each in Proteobacteria, Thermodesulfobacteria, Nitrospirae, Verrucomicrobia, and Planctomycetes, as well as in some unclassified bacteria (Nickel et al. 2017; Daniels et al. 2019). Whereas HARP-encoding bacteria that lack the genes for RNA-based RNase P are found in the Aquificae, Nitrospirae and among the aforementioned unclassified bacteria, HARP as well as *rnpA* and *rnpB* genes are identifiable in all other cases. Larger numbers of HARP genes were found in archaeal genomes, many in Euryarchaeota, some in the TACK, but none in the DPANN and Asgard superphyla; however, all of these archaea encoding a HARP also harbor the RNA and protein genes for RNP RNase P (Nickel et al. 2017; Daniels et al. 2019). It has been noticed that HARP distribution among Euryarchaeota is not only patchy at the class level but often also extends to the genus level (e.g., among Halobacteria or Thermoprotei) (Daniels et al. 2019). Overall, these observations suggest that the function of HARP is not that of a housekeeping RNase P function in Archaea. This view is supported by our recent findings that HARP gene knockouts in the two Euryarchaeota *Haloferax volcanii* and *Methanosarcina mazei* did not result in growth phenotypes under standard conditions, temperature and salt stress (*H. volcanii*) or nitrogen deficiency (*M. mazei*) (Schwarz et al. 2019), whereas attempts to delete the RNase P RNA gene in *H. volcanii* were unsuccessful and its knockdown to ~20% of the wild-type RNase P RNA level impaired tRNA processing and caused retarded cell growth (Stachler and Marchfelder 2016). The patchy occurrence of HARP among archaea could thus be explained by sporadic losses of the gene owing to its non-essentiality.

How did the Aquificaceae acquire their HARP gene? *A. aeolicus* was estimated to have acquired at least ~10% of its protein-coding genes by horizontal gene transfer from archaeal organisms (Aravind et al. 1998). Combined with the fact that HARPs are more frequently found among archaea including many thermophilic Euryarchaeota, but sparsely in Bacteria, might suggest the possibility that the common ancestor of Aquificaceae acquired the HARP gene from an archaeon by horizontal gene transfer. However, this remains speculative at present. A maximum likelihood phylogenetic tree grouped the bacterial HARPs separately from the archaeal ones with a reliable bootstrap value of 0.95 (Nickel et al. 2017). This might be taken as evidence that HARP arose before the separation of Bacteria and Archaea, but was lost from the majority of bacteria and many archaea. An exception is the HARP of Thermococci (Archaea) that groups with the bacterial HARPs, which may point to horizontal gene transfer events.

We demonstrated RNase P activity for several archaeal HARPs as well as for the HARP of a bacterium that also encodes *rnpA* and *rnpB* genes (Nickel et al. 2017; Schwarz et al. 2019). This activity was tested in RNase P processing assays in vitro or via complementation of an *E. coli* strain with repressible expression of its endogenous RNase P. In vitro processing and in vivo complementation activities varied in efficiency, but all tested HARPs showed some RNase P activity in at least one of the two assays. We conclude that HARPs, even if they are not the major housekeeping RNase P activity in the respective organisms, nevertheless have the

basic enzymatic capacity to recognize pre-tRNA substrates and to carry out the RNase P-specific phosphodiester hydrolysis reaction. This makes an evolutionary RNase P replacement scenario in the ancestor of Aquificaceae plausible where a HARP protein had to be reshaped and optimized with only modest effort to become a “full-time” RNase P of the organism. At the same time, this brings up the question as to the biological function of HARPs in Archaea, all of which also contain an RNA-based RNase P. It has not escaped notice that about one-fifth of bacterial and archaeal proteins in the PIN domain-like superfamily are components of toxin–antitoxin systems (Daniels et al. 2019; Gobert et al. 2019; Schwarz et al. 2019) and HARPs were classed with the VapC group (Matelska et al. 2017), which refers to bacterial and archaeal VapC protein toxins that interfere with translation by cleaving mRNAs, ribosomal RNAs (rRNAs) or tRNAs (reviewed in Senissar et al. 2017; Gobert et al. 2019; Schwarz et al. 2019). If HARPs function in this direction, it is puzzling why they carry out tRNA 5'-end maturation instead of inactivating tRNAs via cleavage, e.g., in the anticodon loop as expected for such kind of toxin. Another possibility is that HARP functions as a backup RNase P activity under certain stress conditions where expression and assembly of the RNP enzyme are disturbed.

Finally, the question remains why HARP has supplanted the ancient RNP enzyme in bacteria of the family Aquificaceae. One issue is the high growth temperature of these bacteria. One may argue that folding of P RNA, including coordination of catalytic metal ions, and RNP assembly reach their limits at temperatures of up to 95 °C. However, there are examples of RNP RNase P enzymes in hyperthermophiles such as *T. maritima*, a bacterium that can grow at temperatures of up to 90 °C (Huber et al. 1986; Wang et al. 2012) and remains motile up to 105 °C (Gluch et al. 1995). Another trace could be the condensed size of Aquificaceae genomes, suggesting constraints to streamline the genetic repertoire as well as regulatory and metabolic processes. However, the saving of genome space upon substituting *harp* for *rnpA/rnpB* is yet rather negligible. *A. aeolicus* has two 16S-23S-5S rRNA operons, deleting one would save much more space. Also, *A. aeolicus* expresses other non-coding RNAs, such as the signal recognition particle RNA (SRP RNA), tmRNA and 6S RNA, a regulator of RNA polymerase, which is not present in all bacteria (Wehner et al. 2014; Lechner et al. 2014) and thus not essential. These non-coding RNA examples demonstrate that the expression of functional RNPs beyond the ribosome is feasible in the hyperthermophilic bacterium. To finalize this rather speculative discussion, one may argue that a protein-only HARP can exert the RNase P function just as well as the RNP enzyme and its biogenesis is certainly simpler than expressing an RNA and protein subunit that have to be coordinately expressed and assembled to a functional RNP. Thus, according to the motto “Opportunity makes a thief,” the progenitor of Aquificaceae might have had the chance to recruit a HARP and took the chance.

## References

- Adamski J, Jakob FJ (2001) A guide to 17 $\beta$ -hydroxysteroid dehydrogenases. *Mol Cell Endocrinol* 171:1–4
- Altman S, Kirsebom LA (1999) Ribonuclease P. In: Gesteland RF, Cech T, Atkins JF (eds) *The RNA world*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., pp 351–380
- Amero CD, Boomershine WP, Xu Y, Foster M (2008) Solution structure of *Pyrococcus furiosus* RPP21, a component of the archaeal RNase P holoenzyme, and interactions with its RPP29 protein partner. *Biochemistry* 47:11704–11710
- Anantharaman V, Aravind L (2006) The NYN domains: novel predicted RNAses with a PIN domain-like fold. *RNA Biol* 3:18–27
- Anantharaman V, Koonin EV, Aravind L (2002) SPOUT: a class of methyltransferases that includes spoU and trmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases. *J Mol Microbiol Biotechnol* 4:71–75
- Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet* 14:442–444
- Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet* 8:e1002910
- Baum M, Cordier A, Schön A (1996) RNase P from a photosynthetic organelle contains an RNA homologous to the cyanobacterial counterpart. *J Mol Biol* 257:43–52
- Bernal-Bayard P, Puerto-Galán L, Vioque A (2014) RNase P RNA from the recently evolved plastid of *Paulinella* and from algae. *Int J Mol Sci* 15:20859–20875
- Bonnard G, Gobert A, Arrivé M, Pinker F, Salinas-Giegé T, Giegé P (2016) Transfer RNA maturation in *Chlamydomonas* mitochondria, chloroplast and the nucleus by a single RNase P protein. *Plant J* 87:270–280
- Boomershine WP, McElroy CA, Tsai HY, Wilson RC, Gopalan V, Foster MP (2003) Structure of Mth11/Mth Rpp29, an essential protein subunit of archaeal and eukaryotic RNase P. *Proc Natl Acad Sci USA* 100:15398–15403
- Bothwell AL, Stark BC, Altman S (1976) Ribonuclease P substrate specificity: cleavage of a bacteriophage phi80-induced RNA. *Proc Natl Acad Sci USA* 73:1912–1916
- Brillante N, Gößringer M, Lindenhofer D, Toth U, Rossmann W, Hartmann RK (2016) Substrate recognition and cleavage-site selection by a single-subunit protein-only RNase P. *Nucleic Acids Res* 44:2323–2336
- Brown JW, Nolan JM, Haas ES, Rubio MA, Major F, Pace NR (1996) Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc Natl Acad Sci USA* 93:3001–3006
- Buck AH, Kazantsev AV, Dalby AB, Pace NR (2005) Structural perspective on the activation of RNase P RNA by protein. *Nat Struct Mol Biol* 12:958–964
- Burger G, Gray MW, Forget L, Lang BF (2013) Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol* 5:418–438
- Carrara G, Calandra P, Fruscoloni P, Tocchini-Valentini GP (1995) Two helices plus a linker: a small model substrate for eukaryotic RNase P. *Proc Natl Sci USA* 92:2627–2631
- Chamberlain JR, Lee Y, Lane WS, Engelke DR (1998) Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. *Genes Dev* 12:1678–1690
- Chatfield KC, Coughlin CR 2nd, Friederich MW, Gallagher RC, Hesselberth JR, Lovell MA, Ofman R, Swanson MA, Thomas JA, Wanders RJA et al (2015) Mitochondrial energy failure in HSD10 disease is due to defective mtDNA transcript processing. *Mitochondrion* 21:1–10
- Chen JL, Pace NR (1997) Identification of the universally conserved core of ribonuclease P RNA. *RNA* 3:557–560
- Chen WY, Pulkunat DK, Cho IM, Tsai HY, Gopalan V (2010) Dissecting functional cooperation among protein subunits in archaeal RNase P, a catalytic ribonucleoprotein complex. *Nucleic Acids Res* 38:8316–8327

- Clark WC, Evans ME, Dominissini D, Zheng G, Pan T (2016) tRNA base methylation identification and quantification via high-throughput sequencing. *RNA* 22:1771–1784
- Cordier A, Schön A (1999) Cyanelle RNase P: RNA structure analysis and holoenzyme properties of an organellar ribonucleoprotein enzyme. *J Mol Biol* 289:9–20
- Coughlin DJ, Pleiss JA, Walker SC, Whitworth GB, Engelke DR (2008) Genome-wide search for yeast RNase P substrates reveals role in maturation of intron-encoded box C/D small nucleolar RNAs. *Proc Natl Acad Sci USA* 105:12218–12223
- Crary SM, Niranjanakumari S, Fierke CA (1998) The protein component of *Bacillus subtilis* ribonuclease P increases catalytic efficiency by enhancing interactions with the 5' leader sequence of pre-tRNA(Asp). *Biochemistry* 37:9409–9416
- Daniels CJ, Lai LB, Chen TH, Gopalan V (2019) Both kinds of RNase P in all domains of life: surprise galore. *RNA* 25:286–291
- Daoud R, Forget L, Lang BF (2012) Yeast mitochondrial RNase P, RNase Z and the RNA degradosome are part of a stable supercomplex. *Nucleic Acids Res* 40:1728–1736
- Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M et al (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 392:353–358
- Deutschmann AJ, Amberger A, Zavadil C, Steinbeisser H, Mayr JA, Feichtinger RG, Oerum S, Yue WW, Zschocke J (2014) Mutation or knock-down of 17 $\beta$ -hydroxysteroid dehydrogenase type 10 cause loss of MRPP1 and impaired processing of mitochondrial heavy strand transcripts. *Hum Mol Genet* 23:3618–3628
- Eder PS, Kekuda R, Stolc V, Altman S (1997) Characterization of two scleroderma autoimmune antigens that copurify with human ribonuclease P. *Proc Natl Acad Sci USA* 94:1101–1106
- Engelke DR, Fierke CA (2015) The evolution of RNase P. *RNA* 21:517–518
- Esakova O, Krasilnikov AS (2010) Of proteins and RNA: The RNase P/MRP family. *RNA* 16:1725–1747
- Esakova O, Perederina A, Quan C, Schmitt ME, Krasilnikov AS (2008) Footprinting analysis demonstrates extensive similarity between eukaryotic RNase P and RNase MRP holoenzymes. *RNA* 14:1558–1567
- Evans D, Marquez SM, Pace NR (2006) RNase P: interface of the RNA and protein worlds. *Trends Biochem Sci* 31:333–341
- Forti F, Sabbattini P, Sironi G, Zangrossi S, Deho G, Ghisotti D (1995) Immunity determinant of phage-plasmid P4 is a short processed RNA. *J Mol Biol* 249:869–878
- Fukuhara H, Kifusa M, Watanabe M, Terada A, Honda T, Numata T, Kakuta Y, Kimura M (2006) A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from *Pyrococcus horikoshii* OT3. *Biochem Biophys Res Commun* 343:956–964
- Frank DN, Harris ME, Pace NR (1994) Rational design of self-cleaving pre-tRNA-ribonuclease P RNA conjugates. *Biochemistry* 33:10800–10808
- Frank DN, Adamidi C, Ehringer MA, Pitulle C, Pace NR (2000) Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. *RNA* 6:1895–1904
- Garcia PD, Leach RW, Wadsworth GM, Choudhary K, Li H, Aviran S, Kim HD, Zakian VA (2020) Stability and nuclear localization of yeast telomerase depend on protein components of RNase P/MRP. *Nat Commun* 11:2173
- Giegé R, Jühling F, Pütz J, Stadler P, Sauter C, Florentz C (2012) Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip Rev RNA* 3:37–61
- Gluch ME, Typke D, Baumeister W (1995) Motility and thermotactic responses of *Thermotoga maritima*. *J Bacteriol* 177:5473–5479
- Gobert A, Gutmann B, Taschner A, Gößringer M, Holzmann J, Hartmann RK, Rossmannith W, Giegé P (2010) A single Arabidopsis organellar protein has RNase P activity. *Nat Struct Mol Biol* 17:740–744
- Gobert A, Pinker F, Fuchsbaauer O, Gutmann B, Boutin R, Roblin P, Sauter C, Giegé P (2013) Structural insights into protein-only RNase P complexed with tRNA. *Nat Commun* 4:1353

- Gobert A, Bruggemann M, Giegé P (2019) Involvement of PIN-like domain nucleases in tRNA processing and translation regulation. *IUBMB Life* 71:1117–1125
- Gösringer M, Hartmann RK (2007) Function of heterologous and truncated RNase P proteins in *Bacillus subtilis*. *Mol Microbiol* 66:801–813
- Gösringer M, Lechner M, Brillante N, Weber C, Rossmann W, Hartmann RK (2017) Protein-only RNase P function in *Escherichia coli*: viability, processing defects and differences between PRORP isoenzymes. *Nucleic Acids Res* 45:7441–7454
- Gösringer M, Schencking I, Hartmann RK (2020) RNase P ribozymes. In: Müller S, Winkler W (eds) *Ribozymes*. Wiley-VCH, Weinheim, Germany
- Gray MW, Gopalan V (2020) Piece by piece: building a ribozyme. *J Biol Chem* 295:2313–2323
- Green CJ, Rivera-León R, Vold BS (1996) The catalytic core of RNase P. *Nucleic Acids Res* 24:1497–1503
- Guerrier-Takada C, Gardiner K, Marsh T, Pace NR, Altman S (1983) The RNA moiety of Ribonuclease P is the catalytic subunit of the enzyme. *Cell* 35:849–857
- Guo X, Campbell FE, Sun L, Christian EL, Anderson VE, Harris ME (2006) RNA-dependent folding and stabilization of C5 protein during assembly of the *E. Coli* RNase P holoenzyme. *J Mol Biol* 360:190–203
- Gupta RS, Lali R (2013) Molecular signatures for the phylum aquificae and its different clades: proposal for division of the phylum aquificae into the emended order Aquificales, containing the families aquificaceae and hydrogenothermaceae, and a new order desulfurobacteriales Ord. Nov., containing the family desulfurobacteriaceae. *Antonie Van Leeuwenhoek* 104:349–368
- Gutmann B, Gobert A, Giegé P (2012) PRORP proteins support RNase P activity in both organelles and the nucleus in Arabidopsis. *Genes Dev* 26:1022–1027
- Haas ES, Brown JW (1998) Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res* 26:4093–4099
- Haas ES, Armbruster DW, Vucson BM, Daniels CJ, Brown JW (1996) Comparative analysis of ribonuclease P RNA structure in archaea. *Nucleic Acids Res* 24:1252–1259
- Hartmann E, Hartmann RK (2003) The enigma of ribonuclease P evolution. *Trends Genet* 19:561–569
- Hartmann RK, Heinrich J, Schlegel J, Schuster H (1995) Precursor of C4 antisense RNA of bacteriophages P1 and P7 is a substrate for RNase P of *Escherichia coli*. *Proc Natl Acad Sci USA* 92:5822–5826
- Hartmann RK, Gösringer M, Späth B, Fischer S, Marchfelder A (2009) The making of tRNAs and more—RNase P and tRNase Z. *Prog Mol Biol Transl Sci* 85:319–368
- Harris JK, Haas ES, Williams D, Frank DN, Brown JW (2001) New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA* 7:220–232
- He X-Y, Merz G, Mehta P, Schulz H, Yang S-Y (1999) Human brain short chain L-3-hydroxyacyl coenzyme A dehydrogenase is a single-domain multifunctional enzyme. Characterization of a novel 17 $\beta$ -hydroxysteroid dehydrogenase. *J Biol Chem* 274:15014–15019
- He X-Y, Merz G, Yang Y-Z, Pullarkat R, Mehta P, Schulz H, Yang S-Y (2000a) Function of human brain short chain L-3-hydroxyacyl coenzyme A dehydrogenase in androgen metabolism. *Biochem Biophys Acta* 1484:267–277
- He X-Y, Yang Y-Z, Schulz H, Yang S-Y (2000b) Intrinsic alcohol dehydrogenase and hydroxysteroid dehydrogenase activities of human mitochondrial short-chain L-3-hydroxyacyl-CoA dehydrogenase. *Biochem J* 345:139–143
- He X-Y, Wegiel J, Yang S-Y (2005a) Intracellular oxidation of allopregnanolone by human brain type 10 17 $\beta$ -hydroxysteroid dehydrogenase. *Brain Res* 1040:29–35
- He X-Y, Wegiel J, Yang Y-Z, Pullarkat R, Schulz H, Yang S-Y (2005b) Type 10 17 $\beta$ -hydroxysteroid dehydrogenase catalyzing the oxidation of steroid modulators of  $\gamma$ -aminobutyric acid type A receptors. *Mol Cell Endocrinol* 229:111–117
- Heide C, Busch S, Feltens R, Hartmann RK (2001) Distinct modes of mature and precursor tRNA binding to *Escherichia coli* RNase P RNA revealed by NAIM analyses. *RNA* 7:553–564

- Hipp K, Galani K, Batische C, Prinz S, Böttcher B (2012) Modular architecture of eukaryotic RNase P and RNase MRP revealed by electron microscopy. *Nucleic Acids Res* 40:3275–3288
- Holzmann J, Frank P, Löffler E, Bennett KL, Gerner C, Rossmanith W (2008) RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. *Cell* 135:462–474
- Honda T, Kakuta Y, Kimura K, Saho J, Kimura M (2008) Structure of an archaeal homolog of the human protein complex Rpp21-Rpp29 that is a key core component for the assembly of active ribonuclease P. *J Mol Biol* 384:652–662
- Howard MJ, Lim WH, Fierke CA, Koutmos M (2012) Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing. *Proc Natl Acad Sci USA* 109:16149–16154
- Howard MJ, Liu X, Lim WH, Klemm BP, Koutmos M, Engelke DR, Fierke CA (2013) RNase P enzymes: divergent scaffolds for a conserved biological reaction. *RNA Biol* 10:909–914
- Howell NW, Jora M, Jepson BF, Limbach PA, Jackman JE (2019) Distinct substrate specificities of the human tRNA methyltransferases TRMT10A and TRMT10B. *RNA* 25:1366–1376
- Hsieh J, Fierke CA (2009) Conformational change in the *Bacillus subtilis* RNase P holoenzyme-pre-tRNA complex enhances substrate affinity and limits cleavage rate. *RNA* 15:1565–1577
- Huber R, Langworthy TA, König H, Thomm M, Woese CR, Sleytr UB, Stetter KO (1986) *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 °C. *Arch Microbiol* 144:324–333
- Jackman JE, Montange RK, Malik HS, Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. *RNA* 9:574–585
- Jarrous N (2002) Human ribonuclease P: subunits, function, and intranuclear localization. *RNA* 8:1–7
- Jarrous N (2017) Roles of RNase P and its subunits. *Trends Genet* 33:594–603
- Jarrous N, Altman S (2001) Human ribonuclease P. *Methods Enzymol* 342:93–100
- Jarrous N, Gopalan V (2010) Archaeal/Eukaryal RNase P: subunits, functions and RNA diversification. *Nucleic Acids Res* 38:7885–7894
- Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* 37:159–162
- Jühling T, Duchardt-Ferner E, Bonin S, Wöhnert J, Pütz J, Florentz C, Betat H, Sauter C, Mörl M (2018) Small but large enough: structural properties of armless mitochondrial tRNAs from the nematode *Romanomermis culicivorax*. *Nucleic Acids Res* 46:9170–9180
- Kakuta Y, Ishimatsu I, Numata T, Kimura K, Yao M, Tanaka I, Kimura M (2005) Crystal structure of a ribonuclease P protein Ph1601p from *Pyrococcus horikoshii* OT3: an archaeal homologue of human nuclear ribonuclease P protein Rpp21. *Biochemistry* 44:12086–12093
- Kallberg Y, Oppermann U, Jörnvall H, Persson B (2002) Short-chain dehydrogenases/reductases (SDRs). *Eur J Biochem* 269:4409–4417
- Karasik A, Fierke CA, Koutmos M (2019) Interplay between substrate recognition, 5' end tRNA processing and methylation activity of human mitochondrial RNase P. *RNA* 25:1646–1660
- Kawano S, Nakashima T, Kakuta Y, Tanaka I, Kimura M (2006) Crystal structure of protein Ph1481p in complex with protein Ph1877p of archaeal RNase P from *Pyrococcus horikoshii* OT3: implication of dimer formation of the holoenzyme. *J Mol Biol* 357:583–591
- Kazantsev AV, Krivenko AA, Harrington DJ, Carter RJ, Holbrook SR, Adams PD, Pace NR (2003) High-resolution structure of RNase P protein from *Thermotoga maritima*. *Proc Natl Acad Sci USA* 100:7497–7502
- Kazantsev AV, Krivenko AA, Harrington DJ, Holbrook SR, Adams PD, Pace NR (2005) Crystal structure of a bacterial ribonuclease P RNA. *Proc Natl Acad Sci USA* 102:13392–13397
- Keeling PJ (2004) Diversity and evolutionary history of plastids and their hosts. *Am J Bot* 91:1481–1493
- Kempnaers M, Roovers M, Oudjama Y, Tkaczuk KL, Bujnicki JM, Droogmans L (2010) New archaeal methyltransferases forming 1-methyladenosine or 1-methyladenosine and 1-methylguanosine at position 9 of tRNA. *Nucleic Acids Res* 38:6533–6543



- Kikovska E, Svärd SG, Kirsebom LA (2007) Eukaryotic RNase P RNA mediates cleavage in the absence of protein. *Proc Natl Acad Sci USA* 104:2062–2067
- Kikovska E, Wu S, Mao G, Kirsebom LA (2012) Cleavage mediated by the P15 domain of bacterial RNase P RNA. *Nucleic Acids Res* 40:2224–2233
- Kikuchi Y, Sasaki-Tozawa N, Suzuki K (1993) Artificial self-cleaving molecules consisting of a tRNA precursor and the catalytic RNA of RNase P. *Nucleic Acids Res* 21:4685–4689
- Kikuchi Y, Suzuki-Fujita K (1995) Synthesis and self-cleavage reaction of a chimeric molecule between RNase P-RNA and its model substrate. *J Biochem* 117:197–200
- Kissinger CR, Rejto PA, Pelletier LA, Thomson JA, Showalter RE, Abreo MA, Agree CS, Margosiak S, Meng JJ, Aust RM et al (2004) Crystal structure of human ABAD/HSD10 with a bound inhibitor: implications for design of Alzheimer's disease therapeutics. *J Mol Biol* 342:943–952
- Klemm BP, Wu N, Chen Y, Liu X, Kaitany KJ, Howard MJ, Fierke CA (2016) The diversity of ribonuclease P: protein and RNA catalysts with analogous biological functions. *Biomolecules* 6:27
- Krasilnikov AS, Yang X, Pan T, Mondragón A (2003) Crystal structure of the specificity domain of ribonuclease P. *Nature* 421:760–764
- Krishnamohan A, Jackman JE (2019) A family divided: distinct structural and mechanistic features of the SpoU-TrmD (SPOUT) methyltransferase superfamily. *Biochemistry* 58:336–345
- Kurz JC, Fierke CA (2002) The affinity of magnesium binding sites in the *Bacillus subtilis* RNase P  $\times$  pre-tRNA complex is enhanced by the protein subunit. *Biochemistry* 41:9545–9558
- Kurz JC, Niranjankumari S, Fierke CA (1998) Protein component of *Bacillus subtilis* RNase P specifically enhances the affinity for precursor-tRNA<sup>Asp</sup>. *Biochemistry* 37:2393–2400
- LaGrandeur TE, Darr SC, Haas ES, Pace NR (1993) Characterization of the RNase P RNA of *Sulfolobus acidocaldarius*. *J Bacteriol* 175:5043–5048
- Lai LB, Chan PP, Cozen AE, Bernick DL, Brown JW, Gopalan V, Lowe TM (2010) Discovery of a minimal form of RNase P in *Pyrobaculum*. *Proc Natl Acad Sci USA* 107:22493–22498
- Lai LB, Bernal-Bayard P, Mohannath G, Lai SM, Gopalan V, Vioque A (2011) A functional RNase P protein subunit of bacterial origin in some eukaryotes. *Mol Genet Genomics* 286:359–369
- Lan P, Tan M, Zhang Y, Niu S, Chen J, Shi S, Qiu S, Wang X, Peng X, Cai G et al (2018) Structural insight into precursor tRNA processing by yeast ribonuclease P. *Science* 362(6415):eaat6678
- Lechner M, Nickel AI, Wehner S, Riege K, Wieseke N, Beckmann BM, Hartmann RK, Marz M (2014) Genomewide comparison and novel ncRNAs of Aquificales. *BMC Genomics* 15:522
- Lechner M, Rossmannith W, Hartmann RK, Thölken C, Gutmann B, Giegé P, Gobert A (2015) Distribution of ribonucleoprotein and protein-only RNase P in Eukarya. *Mol Biol Evol* 32:3186–3193
- Lee YC, Lee BJ, Kang HS (1996) The RNA component of Mitochondrial ribonuclease P from *Aspergillus nidulans*. *Eur J Biochem* 235:297–303
- Li Y, Altman S (2004) In search of RNase P RNA from microbial genomes. *RNA* 10:1533–1540
- Li D, Willkomm DK, Schön A, Hartmann RK (2007) RNase P of the *Cyanophora paradoxa* cyanelle: a plastid ribozyme. *Biochimie* 89:1528–1538
- Li D, Willkomm DK, Hartmann RK (2009) Minor changes Largely restore catalytic activity of archaeal RNase P RNA from methanothermobacter thermoautotrophicus. *Nucleic Acids Res* 37:231–242
- Li D, Gößringer M, Hartmann RK (2011) Archaeal-bacterial chimeric RNase P RNAs: towards understanding RNA's architecture, function and evolution. *ChemBioChem* 12:1536–1543
- Li F, Liu X, Zhou W, Yang X, Shen Y (2015) Auto-inhibitory mechanism of the human mitochondrial RNase P protein complex. *Sci Rep* 5:9878
- Lombo RB, Kaberdin VR (2008) RNA Processing in *Aquifex aeolicus* involves RNase E/G and an RNase P-like activity. *Biochem Biophys Res Commun* 366:457–463
- Lopez Sanchez MIG, Mercer TR, Davies SMK, Shearwood A-MJ, Nygård KKA, Richman TR, Mattick JS, Rackham O, Filipovska A (2011) RNA processing in human mitochondria. *Cell Cycle* 10:2904–2916

- Loria A, Pan T (1999) The cleavage step of ribonuclease P catalysis is determined by ribozyme-substrate interactions both distal and proximal to the cleavage site. *Biochemistry* 38:8612–8620
- Lukeš J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* 63:528–537
- Luo MJ, Mao L-F, Schulz H (1995) Short-chain 3-hydroxy-2-methylacyl-CoA dehydrogenase from rat liver: purification and characterization of a novel enzyme of isoleucine metabolism. *Arch Biochem Biophys* 321:214–220
- Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B et al (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* 16:2089–2103
- Lustbader JW, Cirilli M, Lin C, Xu HW, Takuma K, Wang N, Caspersen C, Chen X, Pollak S, Chaney M et al (2004) ABAD directly links A $\beta$  to mitochondrial toxicity in Alzheimer's disease. *Science* 304:448–452
- Maizels N, Weiner AM (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc Natl Acad Sci USA* 91:6729–6734
- Marquez SM, Harris JK, Kelley ST, Brown JW, Dawson SC, Roberts EC, Pace NR (2005) Structural implications of novel diversity in eucaryal RNase P RNA. *RNA* 11:739–751
- Marszalkowski M, Teune JH, Steger G, Hartmann RK, Willkomm DK (2006) Thermostable RNase P RNAs lacking P18 identified in the Aquificales. *RNA* 12:1915–1921
- Marszalkowski M, Willkomm DK, Hartmann RK (2008a) 5'-end maturation of tRNA in *Aquifex aeolicus*. *Biol Chem* 389:395–403
- Marszalkowski M, Willkomm DK, Hartmann RK (2008b) Structural basis of a ribozyme's thermostability: P1–L9 interdomain interaction in RNase P RNA. *RNA* 14:127–133
- Marvin MC, Engelke DR (2009) RNase P: increased versatility through protein complexity? *RNA Biol* 6:40–42
- Marvin MC, Clauder-Münster S, Walker SC, Sarkeshik A, Yates JR 3rd, Steinmetz LM, Engelke DR (2011) Accumulation of noncoding RNA due to an RNase P defect in *Saccharomyces cerevisiae*. *RNA* 17:1441–1450
- Massire C, Jaeger L, Westhof E (1998) Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J Mol Biol* 279:773–793
- Matelska D, Steczkiewicz K, Ginalski K (2017) Comprehensive classification of the PIN domain-like superfamily. *Nucleic Acids Res* 45:6995–7020
- Morales MJ, Wise CA, Hollingsworth MJ, Martin NC (1989) Characterization of yeast mitochondrial RNase P: an intact RNA subunit is not essential for activity in vitro. *Nucleic Acids Res* 17:6865–6881
- Morales MJ, Dang YL, Lou YC, Sulo P, Martin NC (1992) A 105-kDa protein is required for yeast mitochondrial RNase P activity. *Proc Natl Acad Sci USA* 89:9875–9879
- Nickel AI, Wäber NB, Größinger M, Lechner M, Linne U, Toth U, Rossmann W, Hartmann RK (2017) Minimal and RNA-free RNase P in *Aquifex aeolicus*. *Proc Natl Acad Sci USA* 114:11121–11126
- Nieuwlandt DT, Haas ES, Daniels CJ (1991) The RNA component of RNase P from the archaeobacterium *Haloferax volcanii*. *J Biol Chem* 266:5689–5695
- Niranjanakumari S, Stams T, Crary SM, Christianson DW, Fierke CA (1998) Protein component of the ribozyme ribonuclease P alters substrate recognition by directly contacting precursor tRNA. *Proc Natl Acad Sci USA* 95:15212–15217
- Oerum S, Roovers M, Rambo RP, Kopec J, Bailey HJ, Fitzpatrick F, Newman JA, Newman WG, Amberger A, Zschocke J et al (2018) Structural insight into the human mitochondrial tRNA purine N1-methyltransferase and ribonuclease P complexes. *J Biol Chem* 293:12862–12876
- Ofman R, Ruitter JPN, Feenstra M, Duran M, Poll-The BT, Zschocke J, Ensenauer R, Lehnert W, Sass JO, Sperl W et al (2003) 2-Methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency is caused by mutations in the HADH2 gene. *Am J Hum Genet* 72:1300–1307
- Pannucci JA, Haas ES, Hall TA, Harris JK, Brown JW (1999) RNase P RNAs from some archaea are catalytically active. *Proc Natl Acad Sci USA* 96:7803–7808

- Pascual A, Vioque A (1999) Substrate binding and catalysis by ribonuclease P from cyanobacteria and *Escherichia coli* are affected differently by the 3' terminal CCA in tRNA precursors. *Proc Natl Acad Sci USA* 96:6672–6677
- Perederina A, Esakova O, Koc H, Schmitt ME, Krasilnikov AS (2007) Specific binding of a Pop6/Pop7 heterodimer to the P3 stem of the yeast RNase MRP and RNase P RNAs. *RNA* 13:1648–1655
- Perederina A, Esakova O, Quan C, Khanova E, Krasilnikov AS (2010) Eukaryotic ribonucleases P/MRP: the crystal structure of the P3 domain. *EMBO J* 29:761–769
- Persson B, Kallberg Y, Bray JE, Bruford E, Dellaporta SL, Favia AD, Duarte RG, Jörnvall H, Kavanagh KL, Kedishvili N et al (2009) The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem Biol Interact* 178:94–98
- Pinker F, Schelcher C, Fernandez-Millan P, Gobert A, Birck C, Thureau A, Roblin P, Giegé P, Sauter C (2017) Biophysical analysis of Arabidopsis protein-only RNase P alone and in complex with tRNA provides a refined model of tRNA binding. *J Biol Chem* 292:13904–13913
- Powell AJ, Read JA, Banfield MJ, Gunn-Moore F, Yan SD, Lustbader J, Stern AR, Stern DM, Brady RL (2000) Recognition of structurally diverse substrates by type II 3-hydroxyacyl-CoA dehydrogenase (HADH II)/amyloid- $\beta$  binding alcohol dehydrogenase (ABAD). *J Mol Biol* 303:311–327
- Pulukunat DK, Gopalan V (2008) Studies on *Methanocaldococcus jannaschii* RNase P reveal insights into the roles of RNA and protein cofactors in RNase P catalysis. *Nucleic Acids Res* 36:4172–4180
- Rackham O, Busch JD, Matic S, Siira SJ, Kuznetsova I, Atanassov I, Ermer JA, Shearwood A-MJ, Richman TR, Stewart JB et al (2016) Hierarchical RNA processing is required for mitochondrial ribosome assembly. *Cell Rep* 16:1874–1890
- Randau L, Schröder I, Söll D (2008) Life without RNase P. *Nature* 453:120–123
- Rauschenberger K, Schöler K, Sass JO, Sauer S, Djuric Z, Rumig C, Wolf NI, Okun JG, Kölker S, Schwarz H et al (2010) A non-enzymatic function of 17 $\beta$ -hydroxysteroid dehydrogenase type 10 is required for mitochondrial integrity and cell survival. *EMBO Mol Med* 2:51–62
- Reich C, Olsen GJ, Pace B, Pace NR (1988) Role of the protein moiety of ribonuclease P, a ribonucleoprotein enzyme. *Science* 239:178–181
- Reinhard L, Sridhara S, Hällberg BM (2015) Structure of the nuclease subunit of human mitochondrial RNase P. *Nucleic Acids Res* 43:5664–5672
- Reinhard L, Sridhara S, Hällberg BM (2017) The MRPP1/MRPP2 complex is a tRNA-maturation platform in human mitochondria. *Nucleic Acids Res* 45:12469–12480
- Reiter NJ, Osterman A, Torres-Larios A, Swinger KK, Pan T, Mondragón A (2010) Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature* 468:784–791
- Rosenblad MA, López MD, Piccinelli P, Samuelsson T (2006) Inventory and analysis of the protein subunits of the ribonucleases P and MRP provides further evidence of homology between the yeast and human enzymes. *Nucleic Acids Res* 34:5145–5156
- Rossmannith W, Karwan RM (1998) Characterization of human mitochondrial RNase P: novel aspects in tRNA processing. *Biochem Biophys Res Commun* 247:234–241
- Rossmannith W, Tullo A, Potuschak T, Karwan R, Sbisà E (1995) Human mitochondrial tRNA processing. *J Biol Chem* 270:12885–12891
- Schonauer MS, Kastaniotis AJ, Hiltunen JK, Dieckmann CL (2008) Intersection of RNA processing and the type II fatty acid synthesis pathway in yeast mitochondria. *Mol Cell Biol* 28:6646–6657
- Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends Plant Sci* 13:663–670
- Schwarz TS, Wäber NB, Feyh R, Weidenbach K, Schmitz RA, Marchfelder A, Hartmann RK (2019) Homologs of *Aquifex aeolicus* protein-only RNase P are not the major RNase P activities in the archaea *Haloferax volcanii* and *Methanosarcina mazei*. *IUBMB Life* 71:1109–1116
- Seif ER, Forget L, Martin NC, Lang BF (2003) Mitochondrial RNase P RNAs in ascomycete fungi: lineage-specific variations in RNA secondary structure. *RNA* 9:1073–1083

- Seif E, Cadioux A, Lang BF (2006) Hybrid *E. coli*—mitochondrial ribonuclease P RNAs are catalytically active. *RNA* 12:1661–1670
- Sen A, Karasik A, Shanmuganathan A, Mirkovic E, Koutmos M, Cox RT (2016) Loss of the mitochondrial protein-only ribonuclease P complex causes aberrant tRNA processing and lethality in *Drosophila*. *Nucleic Acids Res* 44:6409–6422
- Senissar M, Manav MC, Brodersen DE (2017) Structural conservation of the PIN domain active site across all domains of life. *Protein Sci* 26:1474–1492
- Shafiqat N, Marschall H-U, Filling C, Nordling E, Wu X-Q, Björk L, Thyberg J, Mårtensson E, Salim S, Jörnvall H et al (2003) Expanded substrate screenings of human and *Drosophila* type 10 17 $\beta$ -hydroxysteroid dehydrogenases (HSDs) reveal multiple specificities in bile acid and steroid hormone metabolism: characterization of multifunctional 3 $\alpha$ /7 $\alpha$ /7 $\beta$ /17 $\beta$ /20 $\beta$ /21-HSD. *Biochem J* 376:49–60
- Sidote DJ, Hoffman DW (2003) NMR structure of an archaeal homologue of ribonuclease P protein Rpp29. *Biochemistry* 42:13541–13550
- Sidote DJ, Heideker J, Hoffman DW (2004) Crystal structure of archaeal ribonuclease P protein aRpp29 from *Archaeoglobus fulgidus*. *Biochemistry* 43:14128–14138
- Siegel RW, Banta AB, Hass ES, Brown JW, Pace NR (1996) *Mycoplasma fermentans* simplifies our view of the catalytic core of ribonuclease P RNA. *RNA* 2:452–462
- Sinapah S, Wu S, Chen Y, Pettersson BMF, Gopalan V, Kirsebom LA (2011) Cleavage of model substrates by archaeal RNase P: role of protein cofactors in cleavage-site selection. *Nucleic Acids Res* 39:1105–1116
- Singh RK, Feller A, Roovers M, Van Elder D, Wauters L, Droogmans L, Versées W (2018) Structural and biochemical analysis of the dual-specificity Trm10 enzyme from *Thermococcus kodakaraensis* prompts reconsideration of its catalytic mechanism. *RNA* 24:1080–1092
- Small ID, Peeters N (2000) The PPR motif—a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem Sci* 25:46–47
- Spitzfaden C, Nicholson N, Jones JJ, Guth S, Lehr R, Prescott CD, Hegg LA, Eggleston DS (2000) The structure of ribonuclease P protein from *Staphylococcus aureus* reveals a unique binding site for single-stranded RNA. *J Mol Biol* 295:105–115
- Stachler AE, Marchfelder A (2016) Gene repression in haloarchaea using the CRISPR (clustered regularly interspaced short palindromic repeats)-Cas I-B system. *J Biol Chem* 291:15226–15242
- Stams T, Niranjankumari S, Fierke CA, Christianson DW (1998) Ribonuclease P protein structure: evolutionary origins in the translational apparatus. *Science* 280:752–755
- Steitz TA, Steitz JA (1993) A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci USA* 90:6498–6502
- Stoltzfus A (1999) On the possibility of constructive neutral evolution. *J Mol Evol* 49:169–181
- Stribinskis V, Ramos KS (2007) Rpm2p, a protein subunit of mitochondrial RNase P, physically and genetically interacts with cytoplasmic processing bodies. *Nucleic Acids Res* 35:1301–1311
- Stribinskis V, Gao GJ, Ellis SR, Martin NC (2001a) Rpm2, the protein subunit of mitochondrial RNase P in *Saccharomyces cerevisiae*, also has a role in the translation of mitochondrially encoded subunits of cytochrome c oxidase. *Genetics* 158:573–585
- Stribinskis V, Gao GJ, Sulo P, Ellis SR, Martin NC (2001b) Rpm2p: separate domains promote tRNA and Rpm1r maturation in *Saccharomyces cerevisiae* mitochondria. *Nucleic Acids Res* 29:3631–3637
- Stribinskis V, Heymann HC, Ellis SR, Steffen MC, Martin NC (2005) Rpm2p, a component of yeast mitochondrial RNase P, acts as a transcriptional activator in the nucleus. *Mol Cell Biol* 25:6546–6558
- Sun L, Harris ME (2007) Evidence that binding of C5 protein to P RNA enhances ribozyme catalysis by influencing active site metal ion affinity. *RNA* 13:1505–1515
- Sun L, Campbell FE, Zahler NH, Harris ME (2006) Evidence that substrate-specific effects of C5 protein lead to uniformity in binding and catalysis by RNase P. *EMBO J* 25:3998–4007

- Suryadi J, Tran EJ, Maxwell ES, Brown BA (2005) The crystal structure of the *Methanocaldococcus jannaschii* multifunctional L7Ae RNA-binding protein reveals an induced-fit interaction with the box C/D RNAs. *Biochemistry* 44:9657–9672
- Suzuki T, Suzuki T (2014) A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. *Nucleic Acids Res* 42:7346–7357
- Suzuki T, Nagao A, Suzuki T (2011) Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. *Annu Rev Genet* 45:299–329
- Swinehart WE, Henderson JC, Jackman JE (2013) Unexpected expansion of tRNA substrate recognition by the yeast m1G9 methyltransferase Trm10. *RNA* 19:1137–1146
- Takagi H, Watanabe M, Kakuta Y, Kamachi R, Numata T, Tanaka I, Kimura M (2004) Crystal structure of the ribonuclease P protein Ph1877p from hyperthermophilic archaeon *Pyrococcus horikoshii* OT3. *Biochem Biophys Res Commun* 319:787–794
- Tallsjö A, Kirsebom LA (1993) Product release is a rate-limiting step during cleavage by the catalytic RNA subunit of *Escherichia Coli* RNase P. *Nucleic Acids Res* 21:51–57
- Taschner A, Weber C, Buzet A, Hartmann RK, Hartig A, Rossmannith W (2012) Nuclear RNase P of *Trypanosoma brucei*: a single protein in place of the multi-component RNA-protein complex. *Cell Reports* 2:19–25
- Teramoto T, Kaltany KJ, Kakuta Y, Kimura M, Fierke CA, Tanaka Hall TM (2020) Pentatripeptide repeats of protein-only RNase P use a distinct mode to recognize conserved bases and structural elements of pre-tRNA. *Nucleic Acids Res*. <https://doi.org/10.1093/nar/gkaa627>
- Thomas BC, Gao L, Stomp D, Li X, Gegenheimer PA (1995) Spinach chloroplast RNase P: a putative protein enzyme. *Nucleic Acids Symp Ser* 1995:95–98
- Tkaczuk KL, Dunin-Horkawicz S, Purta E, Bujnicki JM (2007) Structural and evolutionary bioinformatics of the SPOUT superfamily of methyltransferases. *BMC Bioinf* 8:73
- Tsai HY, Pulukkunat DK, Woznick WK, Gopalan V (2006) Functional reconstitution and characterization of *Pyrococcus furiosus* RNase P. *Proc Natl Acad Sci USA* 103:16147–16152
- Vilardo E, Rossmannith W (2015) Molecular insights into HSD10 disease: impact of SDR5C1 mutations on the human mitochondrial RNase P complex. *Nucleic Acids Res* 43:5112–5119
- Vilardo E, Nachbagauer C, Buzet A, Taschner A, Holzmann J, Rossmannith W (2012) A subcomplex of human mitochondrial RNase P is a bifunctional methyltransferase—extensive moonlighting in mitochondrial tRNA biogenesis. *Nucleic Acids Res* 40:11583–11593
- Vilardo E, Amman F, Toth U, Kotter A, Helm M, Rossmannith W (2020) Functional characterization of the human tRNA methyltransferases TRMT10A and TRMT10B. *Nucleic Acids Res* (in press)
- Wäber NB, Hartmann RK (2019) Aquificae. In: Schmidt TM (ed) *Encyclopedia of microbiology*, 4th edn. Elsevier, UK p, pp 226–233
- Walker SC, Engelke DR (2006) Ribonuclease P: the evolution of an ancient RNA enzyme. *Crit Rev Biochem Mol Biol* 41:77–102
- Walker SC, Marvin MC, Engelke D (2010) Eukaryote RNase P and RNase MRP. In: Liu F, Altman S (eds) *Ribonuclease P*. Springer, New York, pp 173–202
- Wan F, Wang Q, Tan J, Tan M, Chen J, Shi S, Lan P, Wu J, Lei M (2019) Cryo-electron microscopy structure of an archaeal ribonuclease P holoenzyme. *Nat Commun* 10:2617
- Wang MJ, Davis NW, Gegenheimer P (1988) Novel mechanisms for maturation of chloroplast transfer RNA precursors. *EMBO J* 7:1567–1574
- Wang Z, Tong W, Wang Q, Bai X, Chen Z, Zhao J, Xu N, Liu S (2012) The temperature dependent proteomic analysis of *Thermotoga maritima*. *PLoS ONE* 7(10):e46463
- Warnecke JM, Fürste JP, Hardt W-D, Erdmann VA, Hartmann RK (1996) Ribonuclease P (RNase P) RNA is converted to a Cd<sup>2+</sup>-ribozyme by a single Rp-phosphorothioate modification in the precursor tRNA at the RNase P cleavage site. *Proc Natl Acad Sci USA* 93:8924–8928
- Watanabe Y-I, Suematsu T, Ohtsuki T (2014) Losing the stem-loop structure from metazoan mitochondrial tRNAs and co-evolution of interacting factors. *Front Genet* 5:109
- Waugh DS, Pace NR (1990) Complementation of an RNase P RNA (rnpB) gene deletion in *Escherichia coli* by homologous genes from distantly related eubacteria. *J Bacteriol* 172:6316–6322

- Weber C, Hartig A, Hartmann RK, Rossmann W (2014) Playing RNase P evolution: swapping the RNA catalyst for a protein reveals functional uniformity of highly divergent enzyme forms. *PLoS Genet* 10:e1004506
- Wegscheid B, Condon C, Hartmann RK (2006) Type A and B RNase P RNAs are interchangeable in vivo despite substantial biophysical differences. *EMBO Rep* 7:411–417
- Wehner S, Damm K, Hartmann RK, Marz M (2014) Dissemination of 6S RNA among bacteria. *RNA Biol* 11:1467–1478
- Wende S, Platzer EG, Jühling F, Pütz J, Florentz C, Stadler PF, Mörl M (2014) Biological evidence for the world's smallest tRNAs. *Biochimie* 100:151–158
- Willkomm DK, Feltens R, Hartmann RK (2002) tRNA Maturation in *Aquifex aeolicus*. *Biochimie* 84:713–722
- Wilson RC, Bohlen CJ, Foster MP, Bell CE (2006) Structure of Pfu Pop5, an archaeal RNase P protein. *Proc Natl Acad Sci USA* 103:873–878
- Wu J, Niu S, Tan M, Huang C, Li M, Song Y, Wang Q, Chen J, Shi S, Lan P, Lei M (2018) Cryo-EM structure of the human ribonuclease P holoenzyme. *Cell* 175:1393–1404
- Xu Y, Amero CD, Pulkunat DK, Gopalan V, Foster MP (2009) Solution structure of an archaeal RNase P binary protein complex: formation of the 30-kDa complex between *Pyrococcus furiosus* RPP21 and RPP29 is accompanied by coupled protein folding and highlights critical features for protein-protein and protein-RNA. *J Mol Biol* 393:1043–1055
- Yagi Y, Hayashi S, Kobayashi K, Hirayama T, Nakamura T (2013) Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS ONE* 8:e57286
- Yan SD, Fu J, Soto C, Chen X, Zhu H, Al-Mohanna F, Collison K, Zhu A, Stern E, Saido T et al (1997) An intracellular protein that binds amyloid- $\beta$  peptide and mediates neurotoxicity in Alzheimer's disease. *Nature* 389:689–695
- Yan SD, Zhu Y, Stern ED, Hwang YC, Hori O, Ogawa S, Frosch MP, Connolly ES Jr, McTaggart R, Pinsky DJ et al (2000) Amyloid  $\beta$ -peptide-binding alcohol dehydrogenase is a component of the cellular response to nutritional stress. *J Biol Chem* 275:27100–27109
- Yuan Y, Altman S (1995) Substrate recognition by human RNase P: identification of small, model substrates for the enzyme. *EMBO J* 14:159–168
- Ziehler WA, Morris J, Scott FH, Millikin C, Engelke DR (2001) An essential protein-binding domain of nuclear RNase P RNA. *RNA* 7:565–575
- Zschocke J (2012) HSD10 disease: clinical consequences of mutations in the HSD17B10 gene. *J Inher Metab Dis* 35:81–89
- Zschocke J, Ruitter JPN, Brand J, Lindner M, Hoffmann GF, Wanders RJA, Mayatepek E (2000) Progressive infantile neurodegeneration caused by 2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency: a novel inborn error of branched-chain fatty acid and isoleucine metabolism. *Pediatr Res* 48:852–855

# Chapter 12

## An Unusual Evolutionary Strategy: The Origins, Genetic Repertoire, and Implications of Doubly Uniparental Inheritance of Mitochondrial DNA in Bivalves



Donald T. Stewart, Sophie Breton, Emily E. Chase, Brent M. Robicheau, Stefano Bettinazzi, Eric Pante, Noor Youssef, and Manuel A. Garrido-Ramos

**Abstract** Mitochondrial DNA (mtDNA) is typically passed on to progeny only by the female parent. The phenomenon of “doubly uniparental inheritance” (DUI) of mtDNA in many bivalve species is a fascinating exception to the paradigm of strict maternal inheritance of mtDNA. In this review, we survey the current state of knowledge of DUI and discuss several active areas of research in this field. Topics/questions covered include: the number of times DUI evolved (once or multiple origins), the link

---

D. T. Stewart (✉)

Department of Biology, Acadia University, Wolfville, NS B4P 2R6, Canada  
e-mail: [don.stewart@acadiau.ca](mailto:don.stewart@acadiau.ca)

S. Breton · S. Bettinazzi

Département de Sciences Biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada  
e-mail: [s.breton@umontreal.ca](mailto:s.breton@umontreal.ca)

S. Bettinazzi

e-mail: [stefano.bettinazzi@umontreal.ca](mailto:stefano.bettinazzi@umontreal.ca)

E. E. Chase

Institut Méditerranéen d’Océanologie, Aix-Marseille University, 13288 Marseille, France  
e-mail: [emily.chase@mio.osupytheas.fr](mailto:emily.chase@mio.osupytheas.fr)

B. M. Robicheau · N. Youssef

Department of Biology, Dalhousie University, Halifax, NS B3H 4R2, Canada  
e-mail: [brentrobicheau@acadiau.ca](mailto:brentrobicheau@acadiau.ca)

N. Youssef

e-mail: [n.youssef@dal.ca](mailto:n.youssef@dal.ca)

E. Pante

Littoral, Environnement et Sociétés Joint Research Unit, 7266 Centre National de la recherche Scientifique, Université de La Rochelle, La Rochelle, France  
e-mail: [eric.pante@univ-lr.fr](mailto:eric.pante@univ-lr.fr)

M. A. Garrido-Ramos

Departamento de Genética, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain  
e-mail: [mgarrido@ugr.es](mailto:mgarrido@ugr.es)

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,  
[https://doi.org/10.1007/978-3-030-57246-4\\_12](https://doi.org/10.1007/978-3-030-57246-4_12)

301

between DUI and sex determination, the role(s) of mtDNA-encoded non-oxidative phosphorylation genes (i.e. ORFan/*orf* genes) in freshwater mussels, the function of conserved sequence motifs and sperm transmission elements in mtDNA of marine mussels, the challenges of annotating mtDNA genomes of DUI species, the presence of unorthodox features in venerid mtDNA, whether or not *orf* DNA sequences are useful in species-level identification of freshwater mussel, and finally, whether or not there are obvious benefits of DUI. For each topic, we also highlight important avenues for future research within this fascinating field of mitochondrial evolutionary biology.

## 12.1 An Overview of Doubly Uniparental Inheritance of Mitochondrial DNA in Bivalves

Doubly uniparental inheritance (or DUI) of mitochondrial DNA is a highly atypical system of mitochondrial DNA inheritance observed in various bivalves (Breton et al. 2007). Since its discovery (early 1990s; e.g. Fisher and Skibinski 1990; Hoeh et al. 1991; Zouros et al. 1992), the field of DUI research has steadily grown. As summarized in a series of review articles, species that possess DUI *typically* exhibit the following features (but exceptions have been noted in the literature; e.g. Ghiselli et al. 2013; Passamonti and Plazzi 2020): (1) two sex-associated mtDNA lineages [a sperm-transmitted male lineage and an egg-transmitted female-transmitted lineage], (2) females are typically homoplasmic, whereas males are heteroplasmic, (3) male somatic tissue is dominated by female-transmitted [F-type] mtDNA, whereas male gonad, and specifically spermatozoa, contains exclusively male-transmitted [M-type] mtDNA, (4) a relatively fast rate of molecular evolution for both types, but especially M-types, (5) in some lineages of bivalves, the M- and F-types occasionally recombine, giving rise to an F-type mtDNA molecule that is mostly composed of F-type protein/RNA-coding genes, with the insertion of a copy of the control region from the M-type, and (6) the presence of open reading frames with no immediately obvious homology to typical mitochondrial genes associated with the electron transport chain or ATP synthesis (Breton et al. 2007; Passamonti and Ghiselli 2009; Breton et al. 2011a; Zouros 2013; Zouros and Rodakis 2019).

At present, >1200 papers have been published on DUI (based on Google Scholar search for “doubly uniparental inheritance” articles from 1990 onwards). Popular areas of research since the phenomenon was first described in Blue mussels of the genus *Mytilus* have included: the taxonomic distribution of DUI, the rate and pattern of molecular evolution of the F- and M-type genomes, gene complement and genetic features of both F & M genomes in various bivalve orders and families, patterns of tissue specific and intracellular gene expression of the M & F genomes, the role of recombination of mt genomes on efficiency rates of electron transport chain complexes and even sperm swimming speed, as well as many other fundamental biological properties. Nonetheless, many features of this unique molecular



phenomenon remain unknown. Indeed, because DUI differs so fundamentally from the rule of strict maternal inheritance found in all other animals, Zouros (2020), and Zouros and Rodakis (2019) have recently argued that this phenomenon deserves greater attention from the broader cell biology, molecular biology, and genetics communities.

Herein, we review the current state of knowledge in several areas of DUI research and identify what we consider to be the primary areas of future investigations. We focus on the following themes in this review: origin(s) of DUI (did it evolve once or multiple times?), DUI and sex determination, open reading frame (*orf*) genes in freshwater mussel mitochondrial genomes (their role in sex determination/sexual development), conserved DNA sequence motifs and sperm transmission elements in the mitochondrial genome of mytilids, issues in annotating DUI genomes, unorthodox features in venerid mitochondrial genomes, using *f-orf* sequences to improve species-level identification of freshwater mussel species, and finally, we investigate the potential benefits of DUI (in particular, to organismal fitness).

## 12.2 The Origins of Doubly Uniparental Inheritance of Mitochondrial DNA in Bivalves: Did It Evolve Once or Multiple Times?

To our knowledge, doubly uniparental inheritance (DUI) of mitochondrial DNA is only found within the bivalved molluscs, but not all lineages of bivalves exhibit DUI. A recent paper found that DUI has been documented in over 100 species from 12 families of bivalves (Gusman et al. 2016). The number has grown in the past four years (e.g. Pante et al. 2017) and will undoubtedly continue to grow as more of the >9000 known bivalve species are studied (Huber 2010). The lineages of bivalves that exhibit DUI that have been studied most extensively include the marine mussels of the order Mytilida, the marine clams of the order Venerida, and the freshwater mussels of the order Unionida. Some species in the order Nuculanida also exhibit the phenomenon, but recent phylogenetic analyses suggest that these taxa are part of the Mytilida, and not part of the basal bivalve lineage, the Protobranchia (e.g. Gusman et al. 2016).

There are two primary, but not mutually exclusive, hypotheses that have been proposed to explain the origin (or origins) of DUI. The first is that having a separate male-transmitted lineage (that is constrained to be transmitted only through sperm and thus reduce the opportunity for cytoplasmic warfare sensu Hurst and Hoekstra 1994) could allow selection to act on the M-type mtDNA genome in ways that benefit male-associated functions (e.g. sperm swimming speed) (Breton et al. 2007; Jha et al. 2008), while simultaneously avoiding the potentially deleterious effects of having two distinct cytoplasmically transmitted genomes within a species (e.g. Hurst and Hoekstra 1994). The second hypothesis suggests that DUI resulted from a selfish genetic element that (1) invaded a sperm mitochondrial genome and (2) that

managed to survive and become transmitted via sperm thereafter (e.g. endogenization of a selfish viral element in sperm mitochondria; Milani et al. 2014). Indeed, the beneficial effects of having a separate male-transmitted lineage could have been the focus of natural selection following the chance invasion of a selfish genetic element.

The phenomenon of DUI is demonstrably old in some lineages but more difficult to age in others. For example, Hoeh et al. (2002) demonstrated that the M- and F-lineages in freshwater mussels (Unionidae) diverged over 200 MYA (Hoeh et al. 2002). Alternatively, Stewart et al. (2009) argued that because of occasional recombination of M and F genomes and role-reversal of a primarily F-type genome via the M-lineage in certain lineages (such as the marine mussels, family Mytilidae), the true age of DUI in these taxa cannot easily be ascertained. Role-reversal essentially resets the divergence time of the M- and F-type genomes (for all parts of the genome except for the relatively small M-type control region) to 0. There appear to be short DNA sequence motifs in the M-type control regions that play a role in maintaining (and preventing destruction) of these genomes during development such that they safely end up in the developing male gonad and in the spermatozoa (see Kyriakou et al. 2015; Robicheau et al. 2017a, b). Role-reversals have been directly observed in *Mytilus*, and therefore have been inferred in other genera in the family such as *Modiolus* (Robicheau et al. 2017) and *Geukensia* (Lubošny et al. 2020). Because these genera diverged from one another possibly as much as 400 MYA (Lee et al. 2019), the phenomenon is likely extremely old in the bivalves. Furthermore, because role-reversals have been observed in another major order exhibiting the phenomenon, the Venerida (Plazzi et al. 2016), we maintain that the most parsimonious explanation regarding the origin of DUI is that it evolved once early on in the history of the bivalves (e.g. in an ancestor to the autolamellibranchiata), but that recombination and role-reversal events have obscured the true age of DUI as originally suggested by Hoeh et al. (1997).

DUI is, however, absent from some other major lineages of bivalves, such as the Pectinida and Ostrida. Because several hermaphroditic lineages of the family Unionidae have lost DUI (they do not retain a sperm-transmitted M-type; Breton et al. 2011a), and because some species of scallops and oysters in these families exhibit hermaphroditism and extreme sexual plasticity (Collin 2013), we argue that if a Pteriomorph ancestor to the Pectinida and Ostrida passed through a hermaphroditic stage, all descendants of that particular ancestor would not possess DUI, even if some lineages re-evolved gonochorism.

In terms of future research directions in this area, we (as well as others) are exploring the role of sperm transmission elements (STEs; as will be discussed in more detail below) coded for by the M-type genomes that have been shown to interact with proteins from the male gonad but not the female gonad as first described by Kyriakou et al. (2015). We are using in silico approaches to search for evidence of these STEs in all mtDNA genomes of DUI-positive species, which would provide further evidence that DUI evolved once, but is a dynamic system across the diverse lineages of bivalves.

### 12.3 What Is the Link Between DUI and Sex Determination in Bivalves?

Breton et al. (2011a) established a theoretical link between DUI and sex determination or the maintenance of gonochorism (i.e. separate female and male sexes). This was based on the observation that in gonochoric freshwater mussel species, the F- and M-type mt genome lineages are present in females and males, respectively (i.e. these species exhibit DUI), whereas obligate hermaphroditic species lose their M-type mtDNA (Breton et al. 2011a). In addition, an open reading frame (*orf* gene) specific to the F-type lineage (hence the *f-orf*) undergoes marked divergence in hermaphroditic taxa relative to their gonochoristic relatives (These *f-orf* genes, the divergent *h-orf* genes found in hermaphrodites, and the distinct *m-orf* genes found in M-type genomes are discussed below in Sect. 3). Accordingly, DUI may be the first sex determination mechanism or sexual development system in animals that involves the mt genome (Breton et al. 2011a).

Ghiselli et al. (2012) explored the hypothesized link between DUI and sex determination by comparing transcriptomic data of female and male gonads of *Ruditapes philippinarum* (Family Veneridae). They identified over 1500 genes with sex-biased or sex-specific expression. Of male-biased or male-specific genes, several were associated with ubiquitination. Based on known examples of the role of ubiquitin in inheritance of mitochondria, specifically in signalling destruction of paternal mitochondria (Sato and Sato 2013), and sex determination in *Drosophila* (Bayrer et al. 2005) and *C. elegans* (Hodgkin 1987; Hansen and Pilgrim 1999; Starostina et al. 2007; Kulkarni and Smith 2008), Ghiselli et al. (2012) proposed ubiquitin as a promising candidate for a role in DUI, and consequently a connection with sex determination.

Following up on the work of Ghiselli et al. (2012), Milani et al. (2013) analysed the structure and location mRNA transcripts for three genes, *psa*, *birc*, and *anubll* in the Manila clam *R. philippinarum*, which, like *Mytilus*, has sex-biased ratios among progeny. They speculated that these genes, which are homologous to genes involved in ubiquitination in other taxa, play roles in sex determination and could help maintain (or degrade) sperm-derived mitochondria during embryonic development in males (or females, respectively) of DUI species. To pursue the potential link between sex determination mechanisms in bivalves and DUI, Capt et al. (2018) hypothesized that if ubiquitin is responsible for the maintenance or loss of M-type mtDNA in bivalve embryos, then there should be signatures of a homologous ubiquitination process in all species exhibiting DUI. To address this hypothesis, Capt and colleagues produced gonad transcriptomes of the DUI-positive species *Venustaconcha ellipsiformis* and *Utterbackia peninsularis* (Family Unionidae) to compare with previously published transcriptomes from *R. philippinarum* (Ghiselli et al. 2012) and *Hyriopsis schlegelii* (Family Unionidae; Shi et al. 2015). This work permitted the comparison of genes across distantly related species of bivalves possessing DUI. The work of Capt et al. (2018) showed that females possessed 18 differentially expressed ubiquitination genes, and 53 that were male-biased. This included two male-biased ubiquitination genes, which were absent from the female-biased group, *psa6*, a potential factor

in sex determination and/or the sexual maturation of DUI species, and *anubL1*, a potential factor in the tagging of sperm mitochondria to distinguish from female mitochondria. They also included *birc5* and *birc2*, other potential tags that could protect the male mitochondria from degradation. Capt et al. (2018) concluded that a similar set of genes related to sex determination and DUI exists among distantly related species possessing DUI.

Looking forward, Capt et al. (2018) suggested that epigenetic modifications are another interesting area for studying sex determination in bivalves, based on work by Milani et al. 2013. Transcriptomic data in Capt et al. (2019) showed a male-biased expression of DNA methyltransferases (including a DNA methyltransferase associated with mitochondrial, *DNMT1*), histone deacetylases, and histone acetyltransferases, all of which are related to epigenetic modifications. The exploration of the role epigenetic modifications and sex determination in DUI-positive bivalves will progress with further transcriptomic analyses to document genes with male or female-biased expression.

## 12.4 What Do We Know About the *orf* Genes in Freshwater Mussel Mitochondrial Genomes and Their Role in Sex Determination/Sexual Development?

The unique mitochondrial *f*- and *m-orf* genes and their link with the maintenance of gonochorism were discovered in freshwater mussel (FWM) species because of the presence of obligate hermaphrodites within the Unionida (Breton et al. 2011a). Note that occasional hermaphrodites do occur in dominantly gonochoristic species (also true for marine mussels), and among species typically considered obligate hermaphrodites (Ghiselin 1969). Consequently, some hermaphrodites are relatively more recently diverged from their gonochoric ancestors, while some are evolutionarily quite older. For example, the paper pondshell, *Utterbackia imbecillis*, appears to be a relatively recently evolved hermaphroditic species (Riccardi et al. 2019). In theory, obligately hermaphroditic FWM species arise from previously gonochoristic species that have, by some means, produced hermaphrodites (Ghiselin 1969). The main hypothesis is that ancestral gonochoristic individuals/populations could be under selective pressures that favour self-fertilization, particularly if individuals are patchily distributed (Bauer 1987). Habitat modification and introduction of invasive species are two factors that have been identified as contributing to the global decline in freshwater mussels (Lopes-Lima et al. 2014, 2017), and these conditions could lead to an increase in hermaphroditic populations in the future.

Hermaphroditism has evolved independently multiple times among the families Unionidae and Margaritiferidae (Breton et al. 2011a). In each of the hermaphroditic lineages, two patterns are readily apparent: (1) the M-type genome is always lost and (2) the remaining genome, which is now referred to as a hermaphroditic genome, experiences highly divergent evolution in its *orf* gene. The *f-orf* of the F-type

genome of a gonochoric species acquires many nucleotide changes, and thus becomes an *h-orf* within an H-type genome. All *h-orf* genes examined to date have been extremely divergent (Breton et al. 2011a; Chase et al. 2018) exhibiting highly modified hydrophobicity plots of their predicted proteins relative to the predicted proteins of *f-orf* proteins from their closely related sister taxa (Breton et al. 2011; Chase et al. 2018; Stewart et al. submitted). These modifications include additional transmembrane helices (TMHs), repeated segments of DNA, and typically an increase in overall length of the *orf* (Breton et al. 2011a). However, some species' mt genomes possess *h-orfs* that do not adhere to these particular patterns (or even lack an *h-orf* gene), namely *Toxolasma parvum*, *Anodonta cygnea*, and *Anodontites trapesialis* (Chase et al. 2018; Soroka and Burzyński 2017; Guerra et al. 2017). The localization of the *f-orf* protein to locations outside the mitochondria was further documented by Breton et al. (2011a), in which immunoelectron microscopy techniques showed the *f-orf* protein in association with the nucleus as well as the mitochondria. It is crucial to explore expression localization not only for more *f-orfs*, but also *h-orfs* and *m-orfs*. Of most interest would be exploring these *orfs* along the spectrum from female gonochoristic individuals to obligate hermaphrodites. Recent work (Stewart et al. unpublished) has begun this process by using PCR primers to sequence *orfs* in a large number of individuals from a single species (in this case, *Pyganodon cataracta*) from many populations in an attempt to identify populations that may contain nascent hermaphrodites. A focus of future work in this area will be to identify recently formed (or indeed emerging) populations of hermaphroditic freshwater mussels to collect additional data on the patterns noted above. Research questions of particular importance include: (i) whether nascent populations of hermaphroditic mussels lose their M-types immediately, and (ii) the degree of changes occurring over time within the *f-orf/h-orf* regions. Examination of recently hermaphroditized populations/individuals versus older hermaphrodites should provide valuable insight into the functional capacity of *h-orfs*, and the overall role of *orf* regions in DUI and sexual development/sex determination.

We must also point out that exciting work is being conducted on a link between a novel category of RNA molecules called small mitochondrial highly transcribed RNAs or “smithRNAs” by Marco Passamonti’s group in Bologna, Italy. Passamonti and colleagues suggest that these molecules may interact with nuclear gene expression to play a role in sex determination/sexual development. Although currently largely theoretical, this group has proposed that in vivo and in vitro experimentation is now warranted to assess the functionality of smithRNAs in DUI species. For details, see Pozzi et al. (2017).

## 12.5 What Role Do Conserved Sequence Motifs and Sperm Transmission Elements Play in DUI in Mytilid Mussels?

A central tenant of DUI is that the heteroplasmy observed in bivalves is sex-linked (Ghiselli et al. 2019). However, for such a pattern to occur, it logically must follow that genetic mechanism(s) exist to facilitate the cellular recognition of male versus female mitochondria/mitotypes (see Plazzi and Passamonti (2019), and particularly Zouros (2020), for more advanced discussions on genetic signatures/models of DUI). Recent work on the description of sperm transmission elements (STEs) and DUI-associated *orf* genes has attempted to identify major regions presumed to convey female or male genetic signature(s). Studies of *orf* genes have gained considerable interest in the past decade (as discussed in Sect. 3), while the work on STEs is relatively more recent. Burzyński et al. (2003) and Zbawicka et al. (2003) early on indicated that genetic signatures associated with DUI signalling were likely housed in the non-coding portion of mtDNA in species exhibiting DUI, as this was the mtDNA region mainly associated with recombination events that lead to the masculinization (i.e. RM-type formation) of otherwise F-type mitochondrial genomes in mytilids. A region later confirmed as the mtDNA control regions (CR; Cao et al. 2004a, b).

In recent years, Kyriakou et al. (2015) sought to identify paternal mitochondrial signatures associated with DUI through the detailed mapping of a recombined masculinized CR mtDNA. Their analyses showed that a particular portion of the variable domain 1 region or “VD1” of the mitochondrial CR housed a unique portion of sequence conferring a paternal signature within an RM *Mytilus galloprovincialis* under investigation (Kyriakou et al. 2015). Using an Electrophoretic Mobility Shift Assay technique, these authors found that the RM-associated mtDNA segment (which they termed the sperm transmission element or “STE”): forms a protein complex associated with male gonad protein extract (also perinuclear mitochondrial associated), has a unique ~22 bp nucleotide motif, and contains presumed nucleotide folding (Kyriakou et al. 2015). In addition, Kyriakou et al. (2016) showed that the homologous mitochondrial F-derived STE sequence was not able to form a stable protein structure comparable to the M-derived STE. Collectively, this work strongly supports a role for STEs in the cellular identification of *Mytilus* M-types as being “paternally derived” (Kyriakou et al. 2015, 2016).

The discovery of a STE has significantly improved our understanding of DUI genetics (particularly of M-type mtDNA), and a subsequent step for STE research has been to locate STE mtDNA signatures in taxa outside *Mytilis* spp. In Kyriakou et al. (2015), a STE motif is specifically identified in *M. galloprovincialis* and *M. trossulus*, while Robicheau et al. (2018) extended this search by probing in silico with the 22 bp STE motif in both a native and a degenerate (purine or pyrimidine) form. Robicheau et al. (2018) found *Mytilus edulis*, *M. californianus*, *M. modiolus*, and *M. senhousia* to also be harbouring *putative* STE signatures that are relatively less conserved versus the known *Mytilus*-type STE signature in Kyriakou et al. (2015). These exercises have demonstrated that the “unique” and recognizable region of

the *Mytilus* RM-type STE (a mere 22 bp) is inherently difficult to search for in lineages phylogenetically distant from *Mytilus* (e.g. Robicheau et al. 2018). We hypothesize two reasons for this. The first is that the motif as it appears in *Mytilus* may not entirely be suitable for searching in silico within other bivalve species. As an illustration, the level of sequence conservation inferred from Kyriakou et al. (2015) is 86% for an RM- versus M-type of the same *Mytilus* species. As one searches for this motif in taxa that are more and more distantly related to *Mytilus*, one would hypothesize that this low level of conservation (within males of the same species) would only continue to decrease even further, thus making it potentially difficult to locate STEs in silico even in the closest of species. The second reason is that perhaps conventional local alignment methods (namely BLAST; Altschul et al. 1997) may be too stringent to identify small regions of STE similarity. The longest consecutive string of conserved nucleotides in the 86% conserved motif stated above is 7 bp (Kyriakou et al. 2015). The least stringent BLAST parameter “blastn” (for shorter sequences) has a default word size seed of 11 bp to initiate an alignment (note that the word size can be decreased further, but doing so would require foresight) (see McGinnis and Madden 2004). Accordingly, for the novice (even when picking the lowest BLAST stringency parameter with default settings), there is a rather low probability of finding a significant intraspecific STE BLAST hit using the 22 bp motif. A combination of both scenarios (extreme sequence divergence(s) and a rather small nucleotide in silico probe) may also be probable. Given these challenges, more creative means of searching for STEs in silico may be needed moving forward. Future efforts should continue to sequence novel bivalve M-mtDNAs in the aim that additional STEs (or regions uniquely M-derived; either highly conserved or degenerate) may be found.

## 12.6 What Issues Are There in Annotating the mtDNA Genomes of DUI Species?

Metazoan mitochondrial genomes are substantially conserved in terms of genome size, organization and gene content (Boore 1999; Gissi et al. 2008; Breton et al. 2010). Most mitogenomes fall within the 14–16 kb length range, have a stable gene content of 37 intronless genes and have short intergenic regions (Boore 1999; Gissi et al. 2008). The standard gene content includes a pair of genes coding for *12S* and *16S* rRNAs, 22 tRNAs, and a set of 13 genes encoding some of the protein subunits of the mitochondrial respiratory chain complexes and ATP synthase: NADH:ubiquinone oxidoreductase (Complex I; ND1–ND6 and ND4L), ubiquinone: cytochrome c oxidoreductase (Complex III; CYTB), cytochrome c oxidase (Complex IV; COX1–COX3), and ATP synthase (Complex V; ATP6 and ATP8) (Anderson et al. 1981; Garesse and Vallejo 2001; Breton et al. 2010). Additional protein subunits of the mitochondrial machinery, as well as proteins required for mtDNA replication and expression,

are encoded by nuclear genes (Boore 1999; Garesse and Vallejo 2001). Notwithstanding, there are well-documented exceptions to this general view, particularly in invertebrates, including larger genomes than the standard compact mitogenome, genes containing introns and gene gain/losses (Gissi et al. 2008; Breton et al. 2014). In addition, tRNA genes are extremely prone to rearrangement and recruitment (Breton et al. 2014; Plazzi et al. 2016). Particularly, mitochondrial genomes of bivalves are extremely diverse in size (the largest genome is actually that of *Dreissena polymorpha*, the zebra mussel, with ~67 kb; McCartney et al. 2019) and gene arrangement (different mitochondrial gene orderings are unique to different lineages) (Boore et al. 2004; Vallès and Boore 2006; Plazzi et al. 2013, 2016; Guerra et al. 2017). Furthermore, mitogenomes of bivalves, especially those of species with DUI, are characterized by an unusually high rate of evolution, specific non-standard gene content and long intergenic sequences of different sizes collectively called unassigned regions (URs).

Both genes and intergenic sequences of mitogenomes of species with DUI show high rates of sequence evolution. In bivalved molluscs in general, this rate is high and there is a correlation between rearrangement rates and evolutionary rates (Plazzi et al. 2016). Specifically, the *atp6*, *atp8*, *nad2*, *nad4L* and *nad6* protein-coding genes were among the most divergent genes in bivalves (Plazzi et al. 2016). High evolutionary rates not only apply at the interspecific but also at the intraspecific level. F- and M-mtDNAs represent two highly differentiated genomes in species with DUI, a distinctive feature of both their separate evolution during millions of years and the faster evolution of the M genome which experience a more relaxed selective constraint than the F genome (Zouros 2013). Thus, extreme F/M-mtDNA sequence divergence levels have been found in *Modiolus modiolus* (37–40%; Robicheau et al. 2017b), in *Geukensia demissa* (31%, but reaching >50% in the most divergent regions; Lubošný et al. 2020), in *Musculista senhousia* (upwards of 32%; Passamonti et al. 2011) and in freshwater mussels (as much as 52%; Doucet-Beaupré et al. 2010) as well as in the marine clams *R. philippinarum* (34%; Mizi et al. 2005), *Scrobicularia plana* and *Limecola balthica* (~53% between the M and F types for both species; Capt et al. 2020). All of these comparisons exceed the already highly sequence divergence values that were very early on identified in *M. galloprovincialis*, *M. edulis*, and *M. trossulus* (10–20%; Mizi et al. 2005; Breton et al. 2006). Thus, with each bivalve mtDNA genome published, we often learn significantly more about the potential F/M/H sequence divergence thresholds that are possible within and between bivalve species/mitotypes.

The existence of highly divergent sequences might be the cause for difficulties annotating the “missing” *atp8* gene, a rapidly evolving gene in marine mussels, and in bivalves in general (Śmietanka et al. 2010). Bivalves and some species of Nematoda and Platyhelminthes were considered to be lacking *atp8* (Hoffmann et al. 1992; Breton et al. 2010). High rates of sequence divergence lead to the inference that *atp8* was absent in these species, but in fact it was demonstrated that it is present in most bivalves analysed, even in some nematodes (Breton et al. 2010; Plazzi et al. 2016) and that the predicted *atp8* is fully functional in *M. edulis* (Lubosny et al. 2018).



On the contrary to gene losses, DUI species are also characterized by gene acquisitions. For example, the often duplicated *tRNA<sup>Met</sup>* gene in Mytilidae species (Hoffmann et al. 1992; Mizi et al. 2005; Passamonti et al. 2011; Śmietanka et al. 2018) among other cases of tRNA gene duplications in specific lineages (see, for example, Guerra et al. 2017 and Lubośny et al. 2020). More remarkable is the duplicated *cox2* gene in the F mitogenome of the clam *R. philippinarum* or in the M genome of the mytilid *Musculista senhousia*, although their roles are still unknown (Passamonti et al. 2011; Ghiselli et al. 2013; Breton et al. 2014). *Mcox2e* evolved from the extension of the *cox2* gene in the M genome of the freshwater bivalves with DUI (Chakrabarti et al. 2006, 2007; Chapman et al. 2008). A similar extension of the *cox2* gene has also been identified recently in the mytilid *Geukensia demissa* (Lubośny et al. 2020). The *Mcox2e* gene of freshwater bivalves encode a new protein equipped with multiple transmembrane helices that is localized to both inner and outer mitochondrial membranes and might function in male reproduction as a specific label determining the fate of sperm mitochondria in fertilized eggs (Cao et al. 2004a; Cogswell et al. 2006; Chakrabarti et al. 2007). The exciting aspect of gene gains in the mitogenomes of species with DUI resides exactly in the generation of new gene products with functions beyond metabolic roles. Thus, there exist in freshwater mussels an absolute correlation between DUI and the presence of the sex-specific *m-orf* and *f-orf* genes in the M and F mtDNA, respectively, which in turn would be directly involved in the maintenance of gonochorism (see preceding sections) (Breton et al. 2009a, 2011a; Mitchell et al. 2016; Guerra et al. 2017). There is also a specific *f-orf* in marine mussels that has a homolog gene in the F mitogenome of *Musculista senhousia* (Breton et al. 2011b) and codes for a functional protein, although its role in DUI or any other biological process remains to be established (Ouimet et al. 2020). The marine clam *R. philippinarum* also have sex-specific *orfs* that are expressed in the M and F mtDNA (Ghiselli et al. 2013; Milani et al. 2013, 2014). In fact, the RPHM21 protein coded by the *m-orf* of the male-transmitted mtDNA of *R. philippinarum* might be involved in spermatogenesis, reproduction and embryo development (Milani et al. 2014). All these *orfs* were collectively called ORFans (this name was suggested earlier to emphasize that they were additional *orfs* to the standard set of mitochondrial genes but with unknown function). ORFans are among the fastest evolving genes in mitogenomes, and many times it is difficult to align their sequences when they belong to different species (Mitchell et al. 2016; Plazzi et al. 2016; Guerra et al. 2017, 2019). Therefore, there is still a debate about the origin of these ORFans. While the debate of their origin moves between the endogenization of viral genes (Milani et al. 2013, 2014) and the duplication of existing gene or the de novo generation from DNA sequences from the unassigned regions (URs), it has been proposed that the mORF of freshwater mussels evolved from a duplicated and diverged *atp8* gene (Mitchell et al. 2016; Guerra et al. 2017), one of the faster evolving mitochondrial genes (Breton et al. 2010; Śmietanka et al. 2010; Lubosny et al. 2020).

ORFans are located in the intergenic sequences of the mitogenomes of species with DUI. These intergenic sequences are generically called unassigned regions (URs). These URs are extremely variable in number, location, and size, both among

species and within species (*F/M* genomes), but usually possess specific and important features such as tandem repeats and dyad symmetries and, especially the largest ones, might hold the mtDNA control region (Cao et al. 2004b, 2009; Guerra et al. 2014, 2017; Robicheau et al. 2017a). For example, marine bivalves are characterized by control regions that differ between F and M genomes by having the M genomes small motifs called sperm transmission elements (STE) (Kyriakou et al. 2015; Robicheau et al. 2017a). For further discussion of these unorthodox features of the mitochondrial genomes, see Sects. 3, 4 and 6.

## 12.7 Particularly Interesting Cases of Unorthodox Features in Mt Genomes of Venerid Bivalves with DUI

All presently identified freshwater mussel species with DUI possess a 3'-coding extension in their male mitochondrial *cox2* gene (*Mcox2*), that is absent from other animals mtDNAs (Curole and Kocher 2002; Chapman et al. 2008; Doucet-Beaupré et al. 2010; Guerra et al. 2017). This extension is translated and localized in both inner and outer mitochondrial membranes in sperm (Chakrabarti et al. 2006, 2007), and it has been hypothesized that it could play a role in the DUI system (Chapman et al. 2008).

Recently, DUI was discovered in two new species: *S. plana* (Venerida: Semelidae) and *L. balthica* (Venerida: Tellinidae) (Gusman et al. 2016; Pante et al. 2017). Their complete mtDNAs have been published, revealing an intriguing difference of ~10 kb between the F and M mt genomes in both species (Capt et al. 2020). A large part of this difference is due to the exceptionally large size of the *Mcox2* gene, which possesses a >4.5 kb and >3.5 kb insertion in *S. plana* and *L. balthica*, respectively. This insertion is absent in the *Fcox2* gene of both species. In *S. plana*, this in-frame insertion, if translated, means that the *Mcox2* gene would be 5679 bp-long and, to our knowledge, would therefore encode for the longest COX2 protein in the animal kingdom (i.e. 1893 amino acids) (Capt et al. 2020). Although it remains to be determined whether it is indeed transcribed and translated in *S. plana*, it is worth noting that a similar in-frame but shorter insertion of 300 nt has been reported in the DUI species *Meretrix lamarckii* (Venerida: Veneridae) (Bettinazzi et al. 2016).

Quite differently, the *Mcox2* gene in *L. balthica* is split in two by the insertion, which divides the gene into *Mcox2a*, encoding the two transmembrane helices and the “heme-patch” region followed by a complete stop codon (TAA), and *Mcox2b*, encoding an enlarged intermembrane space and the Cu<sub>a</sub> centres (Capt et al. 2020). This situation is confirmed by transcriptomic data (Pante et al. unpublished), which indicate that both regions are transcribed (i.e. with discrete, non-overlapping *Mcox2a* and *Mcox2b* transcripts). As for the insertion in *S. plana*, it remains to be determined whether *Mcox2a* and *Mcox2b* are translated in *L. balthica*. That said, a similar case has been described in another bivalve species, i.e. the freshwater mussel *A. cygnea*, in which a translocation of a portion of the *nad5* gene has been hypothesized to be

transcribed and translated separately from the rest of the gene, with both portions possibly being assembled into a functional NAD5 heterodimer (Chase et al. 2018).

Altogether, the observations about *Mcox2* in *S. plana* and *L. balthica* raise several questions: (1) Is this gigantic *cox2* gene in the male genome (or *Mcox2*) an artefact of the sequenced individuals or a conserved trait in each species? (2) Is this extension expressed or is it rather an intron-like sequence that is spliced? (3) If this sequence is conserved and expressed, how does it evolve relative to the rest of the gene and relative to other typical genes, and is it involved in the mitonuclear network of epistatic interactions leading to OXPHOS functioning? Does it have a function aside from bioenergetics? At the moment, several ongoing research projects attempt to answer these questions but preliminary results obtained from additional *S. plana* male individuals indicate that this insertion is conserved among different individuals from different populations (Tassé et al. unpublished), suggesting that it is most probably functional.

It is possible that these remarkable exceptions from the general mitochondrial pattern in bivalves, such as the extra *f-orf* and *m-orf* genes and the unusual features observed in the *Mcox2* gene, could have adaptive explanations such as the occupation of different environmental niches and/or be related to different breeding systems such as gonochorism or hermaphroditism, or to other intrinsic genetic factors. However, it is not yet understood how and why the mitochondria of bivalves have evolved such different strategies of organizing and transmitting their mt genes. The great stability of gene content, arrangement, structure and inheritance, for more than 550 million years in many metazoan groups, suggests strong stabilizing selection for an optimal animal mt genome. The few animal groups that diverge from this pattern, such as bivalves, are thus of particular utility for investigating the potential causes of this near-universal mt genetic system stability.

## 12.8 Can *F-orf* Sequences Be Used to Improve Species-Level Identification of Freshwater Mussel Species?

The accurate identification of mussels is not only interesting for evolutionary and taxonomic studies but is also necessary for conservation efforts of species at risk or endangered (Zieritz et al. 2010). This is especially relevant for critically imperilled animals such as freshwater mussels (Régnier et al. 2009). Freshwater mussels' vulnerability is commonly linked with anthropogenic factors, including a variety of extrinsic factors such as the loss and change of habitats, pollution, non-native species, and climate change, alongside intrinsic factors of which includes accurate species identification (Ferreira-Rodríguez et al. 2019). As Lopes-Lima et al. (2018) highlighted, there are gaps in our conservation assessments of freshwater mussels globally and these are required for effective conservation actions. Ultimately, the

intrinsic factor of accurate species identification is the foundation of filling these gaps and leading to protection and rehabilitation of species at risk.

Several studies have demonstrated the intraspecific morphological plasticity of freshwater mussels due to environmental factors (e.g. Jerathitikul et al. 2019; Inoue et al. 2013; Reis et al. 2013; Zieritz et al. 2010), and this is a likely factor in species misidentification. Shea et al. (2011) found an average rate of 27% misidentification of Unionidae species in their study and concluded that these rates would significantly impact mussel surveys. Given that conservation actions rely heavily on mussel surveys, we can conclude that misidentification of freshwater mussels could have a downstream effect of eventual misappropriation of conservation actions. An obvious component to solving these issues is the use of DNA barcoding (i.e. molecular markers) for species identification in mussels, which can be used alongside morphological identification for improving the accuracy of mussel surveys. However, recent mussel survey protocols do not include a molecular component (e.g. Smith 2006; Zieritz et al. 2014; a literature review of Freshwater Mussel Survey and Relocation Guidelines Final Report August 2016; West Virginia Mussel Survey Protocols March 2018; Freshwater Mussel Surveys: I-26 Widening Final Report August 2018).

Molecular barcoding provides a more objective way to identify and distinguish between freshwater mussel species (Hebert and Gregory 2005). Species identification by molecular markers often relies on *cox1* (Hebert et al. 2003), *16S* rRNA, and occasionally *nad1* mitochondrial genes. A recent study (Robicheau et al. 2018) assessed mitochondrial protein-coding genes of two recently divergence freshwater mussel taxa and suggested that the female-type ORFan, the *f-orf*, was a useful additional molecular marker for both population and species-level studies of freshwater mussels. A species-level informative marker will evolve enough to detect the differences between two species while being maintained enough to detect individuals within the same species (Robicheau et al. 2018). The addition of *f-orf* sequencing data was shown to improve phylogenetic trees previously based of *cox1* alone in closely related freshwater mussel species. However, here the goal is not solely to improve our phylogenetic hypotheses (which are useful for testing our markers), but instead to improve the data that conservation efforts of freshwater mussels rely on. It is possible that the use of two molecular markers for preliminary work on populations, and/or in combination with large mussel surveys for verification of species identification could reduce errors in population and species reports, and subsequently correctly identify prime targets of conservation actions. Alternatively, the *f-orf* sequence could broadly be used as a primary barcode of life for freshwater mussel species specifically, although much more testing would be required, thereby increasing the effectivity of survey efforts. Ideally, whole mitochondrial genomes would be sequenced, and all individuals would be identified based on these sequences, however, careful planning and identification using both morphological and two molecular markers may help in reducing the significant misidentification of freshwater mussel species in general while maintaining relatively low-cost surveying.

## 12.9 The Potential Benefits of DUI

From an evolutionary perspective, two major fitness issues rise from having a specific genome within mitochondria: (i) the need to preserve the genetic information carried by the mtDNA from OXPHOS mutagenic by-products, and (ii) the need to maintain compatibility between mtDNA- and nDNA-encoded subunits of respiratory complexes. Interestingly, the solution to these problems might reside in the very existence of anisogamy (i.e. gamete dimorphism), together with mitochondrial bioenergetics and transmission mechanism. Under the “division of labour” hypothesis, sperm would sacrifice their own genetic integrity by exploiting the OXPHOS to sustain motility, while oocytes would limit their aerobic metabolism in order to preserve their genetic integrity in quiescent mitochondria (Allen 1996). Then, the strict maternal inheritance of mitochondria in metazoans would retain the solely oocyte-derived mitochondria to be passed on to the next generations, altogether promoting genetic integrity and the avoidance of genetic conflicts through homoplasmy (i.e. a condition in which all mtDNA copies are almost identical). Although beneficial, the selective elimination of sperm mitochondria excludes males from taking part in the evolution of the mitochondrial genome. As a result, deleterious mutations for males can accumulate in the mitochondrial genome and spread in a population just by being neutral or beneficial in females. This sex-specific selective sieve in the evolution of the mitochondrial genome is known as the “mother’s curse” (Gemmell et al. 2004).

Unlike in most animals, male and female mitochondria in DUI species contribute *together* to the genetic pool of progeny. The preservation (and transmission) of a sperm-specific M-type mtDNA challenges the “classic” dynamics of mtDNA evolution, refuting both the “division of labour” and “mother’s curse” concepts. On the one hand, it induces the evolutionary challenge to preserve the genetic integrity of both sex-specific templates to be passed on to the next generation. On the other hand, the very existence of the DUI system represents an unprecedented opportunity for the mitochondrial genome to evolve adaptively for male functions, breaking the female-driven constraints in its evolution. The strict sex-specific mtDNA segregation in DUI germline cause different selective pressures to act on the two mtDNA F & M variants, leading them to evolve distinctly for female and male functions. Sexually antagonist selection is also supported by the extreme divergence between the F- and the M-mtDNA, ranging from 8 to 50% of nucleotide divergence, depending on the gene and species involved (Breton et al. 2007; Passamonti and Ghiselli 2009; Zouros 2013; Capt et al. 2020). Given the exclusive presence of the M-type mtDNA in DUI sperm, and that mutations in the mtDNA are likely to affect the functioning of mitochondrial respiration, a rational indication is that the evolution of DUI male mitochondria could embrace bioenergetic adaptations for male-associated functions, fostering sperm performance and reproductive success (Burt and Trivers 2006; Breton et al. 2007, 2009b).

During the last decade, several studies have investigated the potential benefits of carrying male-derived mitochondria for bivalve sperm performance (Everett et al. 2004; Jha et al. 2008; Stewart et al. 2012; Bettinazzi et al. 2020). Everett et al.

(2004) and Jha et al. (2008) tested this assumption in the DUI species *M. edulis*, comparing motility traits between “standard” sperm, carrying the M-type mtDNA, and “masculinized” sperm, carrying primarily F-type mtDNA with a segment of M-type control region sequence (called RM-types). Interestingly, bearing the M-type mitochondria did not provide any visible advantage in term of speed, with “masculinized” F-type sperm swimming equal or even faster than M-type sperm (Everett et al. 2004; Jha et al. 2008). Congruent results were recently found by Bettinazzi et al. (2020), who extended this investigation to various bivalves with a different mode of mitochondria transmission (i.e. SMI, whose sperm carry the maternal mitochondria, and DUI, whose sperm bear the paternal mitochondria). In sharp contrast with sperm carrying maternally inherited mitochondria (SMI system), M-type sperm of the DUI species *M. edulis* and *R. philippinarum* exhibits a convergent readapted phenotype, characterized by lower speed and accentuate curvilinear trajectory (Bettinazzi et al. 2020). Differences in sperm performance also reflect change in the energetic strategy that fuels sperm motility. For bivalve sperm bearing maternally derived mitochondria (SMI), both aerobic and anaerobic mechanisms of ATP production concur to sustain motility. Conversely, sperm of the DUI species *M. edulis* and *R. philippinarum* appear to strictly rely on OXPHOS in absence of oocytes, but partially switch towards a more combined strategy, implying also fermentation, once detecting egg-derived chemical cues (Bettinazzi et al. 2020). Although a concomitant change in sperm performance is controversial in these very species (Stewart et al. 2012; Bettinazzi et al. 2020), evidence exists that sperm of *M. galloprovincialis* does begin to swim faster and straighter towards eggs that are most genetically similar at the level of the mtDNA, but least similar at the nuclear level (Oliver and Evans 2014; Lymbery et al. 2017). Interestingly, this mechanism could potentially foster heterozygosity and cytonuclear compatibility in offspring, two conditions of likely utmost importance for heteroplasmic DUI species. Overall, evidences support the intriguing hypothesis that the DUI system might indeed be beneficial for sperm performance and fertilization success, at least in the species tested so far. Selection for male functions appears to promote sperm swimming in a slower, more curvilinear and strictly aerobic fashion, in absence of eggs, with the ability to undergo change in performance and bioenergetics once detecting genetically compatible eggs. These traits potentially represent an advantage for the fertilization strategy of broadcast spawning invertebrates, and the DUI system appears to exploit them, potentially enhancing endurance, survival and area covered by sperm. This would in turn maximize the chances of encountering compatible eggs, altogether promoting fertilization success and potential mitonuclear compatibility in open and turbulent marine environments (Levitan 2000; Liu et al. 2011; Fitzpatrick et al. 2012).

The evolutionary consequence of carrying two divergent mitotypes is also evident at the level of cellular and mitochondrial metabolism. First, in opposite trend to SMI species, DUI sperm exhibit a general downregulation of cellular bioenergetics when compared to oocytes. This is evident in the efficiency of key enzymes of glycolysis, fermentation, tricarboxylic acid cycle, fatty acid metabolism and even

OXPPOS (Bettinazzi et al. unpublished results). At the level of mitochondrial functionality, recent findings revealed difference in the functional properties of mitochondria bearing either the paternally or the maternally associated mtDNA. For the SMI species tested, female-transmitted mitochondria exhibit a conserved OXPPOS organization in both gametes and somatic tissues. Conversely, for the DUI species *Arctica islandica* and *M. edulis*, they express convergent OXPPOS remodelling in sperm mitochondria. M-type mitochondria in sperm and male somatic heteroplasmic tissues show functional divergence in OXPPOS activity and organization compared to F-type mitochondria present in eggs and female somatic tissues, involving a strong limitation of the electron transport system (ETS) by the phosphorylation system and a minimal spare capacity at cytochrome *c* oxidase (Bettinazzi et al. 2019). Congruently, a lower activity of cytochrome *c* oxidase (CCO) was also detected for *M. edulis* M-type sperm, in comparison with the F-type “masculinized” ones (Breton et al. 2009b). The existence of a specific DUI mitochondrial remodelling, together with the fact that DUI sperm energetic strategy tightly relies on mitochondrial respiration, supports the expectation that the selective forces driving the evolution of sperm mitochondria in absence of SMI might affect mt encoded components of respiratory complexes, thus foster change in the OXPPOS mechanisms and organization (Breton et al. 2007, 2009b). The over-described architecture provides unusual respiratory control at the terminus of the respiratory chain, high sensitivity to oxygen content in the medium, high ROS flux and, interestingly, the capacity to preserve high membrane potential in sperm mitochondria (Bettinazzi et al. 2019). Accumulating evidence supports this idea that DUI male mitochondria possess the ability to preserve a high mitochondrial membrane potential ( $\Delta\psi_m$ ). This comes from the direct observation of DUI gametes  $\Delta\psi_m$  (Milani and Ghiselli 2015) and from bioenergetic properties of sperm mitochondria in line with the maintenance of a high electrochemical gradient (Bettinazzi et al. 2019, 2020). As the  $\Delta\psi_m$  depicts healthy mitochondria, it has been proposed that the ability to maintain it might play a key role in DUI paternal mitochondria preservation and transmission (Milani 2015).

The findings described here provide evidence of a robust link between the mitochondrial genotype and phenotype in DUI species. Specifically, direct selection on DUI paternally derived mitochondria potentially produces: (i) a widespread down-regulation of cellular bioenergetics, detected at the level of all main energy producing pathways; (ii) the expression of a male-specific mitochondrial phenotype in sperm and partly in heteroplasmic soma, an OXPPOS remodelling providing unusual respiratory control at the terminus of the respiratory chain; (iii) the exhibition of a specific sperm phenotype, characterized by different performance and energetic strategy adopted, which could enhance fertilization success and mitonuclear compatibility; and (iv) the potential ability to preserve a high mitochondrial membrane potential. Overall, these findings suggest that the adaptive value of sex-specific mtDNA variants in DUI could altogether embrace paternal mitochondria preservation, male-specific energetic adaptation, and fertilization success.

**Acknowledgements** DTS and SB were funded by NSERC Discovery grants. EEC was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No713750, carried out with the financial support of the Regional Council of Provence- Alpes-Côte d'Azur and with the financial support of the A\*MIDEX (n° ANR-11-IDEX-0001-02), funded by the Investissements d'Avenir project funded by the French Government, managed by the French National Research Agency [ANR]). At time of submission, an NSERC CGS-D supported BMR. MAGR was supported by a Harrison McCain Visiting Professorship Award from the Harrison McCain Foundation.

## References

- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290:457–465
- Allan JF (1996) Separate sexes and the mitochondrial theory of ageing. *J Theor Biol* 180:135–140
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25:3389–3402
- Bauer G (1987) Reproductive strategy of the freshwater pearl mussel *Margaritifera margaritifera*. *J Anim Ecol* 56:691–704
- Bayrer JR, Zhang W, Weiss MA (2005) Dimerization of doublesex is mediated by a cryptic ubiquitin-associated domain fold: implications for sex-specific gene regulation. *J Biol Chem* 280:32989–32996
- Bettinazzi S, Plazzi F, Passamonti M (2016) The complete female-and male-transmitted mitochondrial genome of *Meretrix lamarckii*. *PLoS one* 11
- Bettinazzi S, Rodríguez E, Milani L, Blier PU, Breton S (2019) Metabolic remodelling associated with mtDNA: insights into the adaptive value of doubly uniparental inheritance of mitochondria. *Proc Roy Soc B* 286:20182708
- Bettinazzi S, Nadarajah S, Dalpé A, Milani L, Blier PU, Breton S (2020) Linking paternally inherited mtDNA variants and sperm performance. *Phil Trans Roy Soc B* 375:20190177
- Boore JL (1999) Animal mitochondrial genomes. *Nucleic Acids Res* 27:1767–1780
- Boore JL, Medina M, Rosenberg LA (2004) Complete sequences of the highly rearranged molluscan mitochondrial genomes of the scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis*. *Mol Biol Evol* 21:1492–1503
- Breton S, Burger G, Stewart DT, Blier PU (2006) Comparative analysis of gender-associated complete mitochondrial genomes in marine mussels (*Mytilus* spp.). *Genetics* 172:1107–1119
- Breton S, Beaupre HD, Stewart DT, Hoeh WR, Blier PU (2007) The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? *Trends Genet* 23:465–474
- Breton S, Beaupré HD, Stewart DT, Piontkivska H, Karmakar M, Bogan AE, Blier PU, Hoeh WR (2009a) Comparative mitochondrial genomics of freshwater mussels (Bivalvia: Unionoida) with doubly uniparental inheritance of mtDNA: gender-specific open reading frames and putative origins of replication. *Genetics* 183:1575–1589
- Breton S, Stewart DT, Blier PU (2009b) Role-reversal of gender-associated mitochondrial DNA affects mitochondrial function in *Mytilus edulis* (Bivalvia: Mytilidae). *J Exp Zool B* 312:108–117
- Breton S, Stewart DT, Hoeh WR (2010) Characterization of a mitochondrial ORF from the gender-associated mtDNAs of *Mytilus* spp. (Bivalvia: Mytilidae): identification of the “missing” ATPase 8 gene. *Mar Genomics* 3:11–18
- Breton S, Stewart DT, Shepardson S, Trdan RJ, Bogan AE, Chapman EG, Ruminas AJ, Piontkivska H, Hoeh WR (2011a) Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? *Mol Biol Evol* 28:1645–1659



- Breton S, Ghiselli F, Passamonti M, Milani L, Stewart DT, Hoeh WR (2011b) Evidence for a fourteenth mtDNA-encoded protein in the female-transmitted mtDNA of marine mussels (Bivalvia: Mytilidae). *PLoS ONE* 6:e19365
- Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M (2014) A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. *Trends Genet* 30:555–564
- Burt A, Trivers R (2006) Selfish mitochondrial DNA. *Genes in Conflict: the biology of selfish genetic elements*. Belknap Press of Harvard University, Cambridge, MA
- Burzyński A, Zbawicka M, Skibinski DOF, Wenne R (2003) Evidence for recombination of mtDNA in the marine mussel *Mytilus trossulus* from the Baltic. *Mol Biol Evol* 20:388–392
- Cao L, Kenchington E, Zouros E (2004a) Differential segregation patterns of sperm mitochondria in embryos of the blue mussel (*Mytilus edulis*). *Genetics* 166:883–894
- Cao L, Kenchington E, Zouros E, Rodakis GC (2004b) Evidence that the large noncoding sequence is the main control region of maternally and paternally transmitted mitochondrial genomes of the marine mussel (*Mytilus* spp.). *Genetics* 167:835–850
- Cao L, Ort BS, Mizi A, Pogson G, Kenchington E, Zouros E, Rodakis GC (2009) The control region of maternally and paternally inherited mitochondrial genomes of three species of the sea mussel genus *Mytilus*. *Genetics* 181:1045–1056
- Capt C, Renaut S, Ghiselli F, Milani L, Johnson NA, Sietman BE, Stewart DT, Breton S (2018) Deciphering the link between doubly uniparental inheritance of mtDNA and sex determination in bivalves: clues from comparative transcriptomics. *Genome Biol Evol* 10:577–590
- Capt C, Renaut S, Stewart DT, Johnson NA, Breton S (2019) Putative mitochondrial sex determination in the Bivalvia: insights from a hybrid transcriptome assembly in freshwater mussels. *Frontiers Genet* 10:840
- Capt C, Bouvet K, Guerra D, Robicheau BM, Stewart DT, Pante E, Breton S (2020) Unorthodox features in two venerid bivalves with doubly uniparental inheritance of mitochondria. *Sci Reports* 10:1–3
- Chakrabarti R, Walker JM, Stewart DT, Trdan RJ, Vijayaraghavana S, Curole JP, Hoeh WR (2006) Presence of a unique male-specific extension of C-terminus to the cytochrome c oxidase subunit II protein coded by the male-transmitted mitochondrial genome of *Venustaconcha ellipsiformis* (Bivalvia: Unionoidea). *FEBS Lett* 580:862–866
- Chakrabarti R, Walker JM, Chapman EG, Shepardson SP, Trdan RJ, Curole JP, Watters GT, Stewart DT, Vijayaraghavana S, Hoeh WR (2007) Reproductive function for a C-terminus extended, male-transmitted cytochrome c oxidase subunit II protein expressed in both spermatozoa and eggs. *FEBS Lett* 581:5213–5219
- Chapman EG, Piontkivska H, Walker JM, Stewart DT, Curole JP, Hoeh WR (2008) Extreme primary and secondary protein structure variability in the chimeric male-transmitted cytochrome c oxidase subunit II protein in freshwater mussels: Evidence for an elevated amino acid substitution rate in the face of domain-specific purifying selection. *BMC Evol Biol* 8:165
- Chase EE, Robicheau BM, Veinot S, Breton S, Stewart DT (2018) The complete mitochondrial genome of the hermaphroditic freshwater mussel *Anodonta cygnea* (Bivalvia: Unionidae): *in silico* analyses of sex-specific ORFs across order Unionoidea. *BMC Genom* 19:221
- Cogswell AT, Kenchington ELR, Zouros E (2006) Segregation of sperm mitochondria in two- and four-cell embryos of the blue mussel *Mytilus edulis*: implications for the mechanism of doubly uniparental inheritance of mitochondrial DNA. *Genome* 49:799–807
- Collin R (2013) Phylogenetic patterns and phenotypic plasticity of molluscan sexual systems. *Integ Comp Biol* 53:723–735
- Curole JP, Kocher TD (2002) Ancient sex-specific extension of the cytochrome c oxidase II gene in bivalves and the fidelity of doubly-uniparental inheritance. *Mol Biol Evol* 19:1323–1328
- Doucet-Beaupré H, Breton S, Chapman EG, Blier PU, Bogan AE, Stewart DT, Hoeh WR (2010) Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. *BMC Evol Biol* 10:50

- Everett EM, Williams PJ, Gibson G, Stewart DT (2004) Mitochondrial DNA polymorphisms and sperm motility in *Mytilus edulis* (Bivalvia: Mytilidae). *J Exp Zool A* 301:906–910
- Ferreira-Rodríguez N, Akiyama YB, Aksenova OV, Araujo R, Barnhart MC, Bespalaya YV, Bogan AE, Bolotov IN, Budha PB, Clavijo C, Clearwater SJ (2019) Research priorities for freshwater mussel conservation assessment. *Biol Conser* 231:77–87
- Fisher C, Skibinski DOF (1990) Sex-biased mitochondrial DNA heteroplasmy in the marine mussel *Mytilus*. *Proc Roy Soc Lond (B)* 242:149–156
- Fitzpatrick JL, Simmons LW, Evans JP (2012) Complex patterns of multivariate selection on the ejaculate of a broadcast spawning marine invertebrate. *Evolution* 66:2451–2460
- Garesse R, Vallejo CG (2001) Animal mitochondrial biogenesis and function: a regulatory cross-talk between two genomes. *Gene* 263:1–16
- Gemmell NJ, Metcalf VJ, Allendorf FW (2004) Mother's curse: the effect of mtDNA on individual fitness and population viability. *Trends Ecol Evol* 19:238–244
- Ghiselin MT (1969) The evolution of hermaphroditism among animals. *Quart Rev Biol* 44:189–208
- Ghiselli F, Maurizii MG, Reunov A, Ariño-Bassols H, Cifaldi C, Pecci A, Alexandrova Y, Bettini S, Passamonti M, Franceschini V, Milani L (2019) Natural heteroplasmy and mitochondrial inheritance in bivalve molluscs. *Integr Comp Biol* 59:1016–1032
- Ghiselli F, Milani L, Chang PL et al (2012) *De novo* assembly of the manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. *Mol Biol Evol* 29:771–786
- Ghiselli F, Milani L, Guerra D, Chang PL, Breton S, Nuzhdin SV, Passamonti M (2013) Structure, transcription, and variability of metazoan mitochondrial genome: perspectives from an unusual mitochondrial inheritance system. *Genome Biol Evol* 5:1535–1554
- Gissi C, Iannelli F, Pesole G (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity* 101:301–320
- Guerra D, Ghiselli F, Passamonti M (2014) The largest unassigned regions of the male and female-transmitted mitochondrial DNAs in *Musculista senhousia* (Bivalvia, Mytilidae). *Gene* 536:316–325
- Guerra D, Lopes-Lima M, Froufe E, Gan HM, Ondina P, Amaro R, Klunzinger MW, Callil C, Prié V, Bogan AE, Stewart DT (2019) Variability of mitochondrial ORFans hints at possible differences in the system of doubly uniparental inheritance of mitochondria among families of freshwater mussels (Bivalvia: Unionida). *BMC Evol Biol* 19
- Guerra D, Plazzi F, Stewart DT, Bogan AE, Hoeh WR, Breton S (2017) Evolution of sex-dependent mtDNA transmission in freshwater mussels (Bivalvia: Unionida). *Sci Rep* 7:1551
- Gusman A, Lecomte S, Stewart DT, Passamonti M, Breton S (2016) Pursuing the quest for better understanding the taxonomic distribution of the system of doubly uniparental inheritance of mtDNA. *PeerJ* 4:e2760
- Hansen D, Pilgrim D (1999) Sex and the single worm: sex determination in the nematode *C. elegans*. *Mech Dev* 83:3–15
- Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Syst Biol* 54:852–859
- Hebert PDN, Ratnasingham S, de Waard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Roy Soc B* 270:S96–S99
- Hodgkin J (1987) A genetic analysis of the sex-determining gene, *tra-1*, in the nematode *Caenorhabditis elegans*. *Genes Dev* 1:731–745
- Hoeh WR, Blakley KH, Brown WM (1991) Heteroplasmy suggests limited biparental inheritance of *Mytilus* mitochondrial DNA. *Science* 251:1488–1490
- Hoeh WR, Stewart DT, Saavedra C, Sutherland BW, Zouros E (1997) Phylogenetic evidence for role-reversals of gender-associated mitochondrial DNA in *Mytilus* (Bivalvia: Mytilidae). *Mol Biol Evol* 14:959–967
- Hoeh WR, Stewart DT, Guttman SI (2002) High fidelity of mitochondrial genome transmission under the doubly uniparental mode of inheritance in freshwater mussels (Bivalvia: Unionoidea). *Evolution* 56:2252–2261

- Hoffmann RJ, Boore JL, Brown WM (1992) A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. *Genetics* 131:397–412
- Huber M (2010) Compendium of bivalves. ConchBooks, Germany
- Hurst LD, Hoekstra RF (1994) Shellfish gees kept in line. *Nature* 368:817–818
- Inoue K, Hayes DM, Harris JL, Christian AD (2013) Phylogenetic and morphometric analyses reveal ecophenotypic plasticity in freshwater mussels *Obovaria jacksoniana* and *Villosa arkansasensis* (Bivalvia: Unionidae). *Ecol Evol* 3:2670–2683
- Jeratthitikul E, Phuangphong S, Sutcharit C et al (2019) Integrative taxonomy reveals phenotypic plasticity in the freshwater mussel *Conradens contradens* (Bivalvia: Unionidae) in Thailand, with a description of a new species. *Syst Biodiv* 17:134–147
- Jha M, Côté J, Hoeh WR, Blier PU, Stewart DT (2008) Sperm motility in *Mytilus edulis* in relation to mitochondrial DNA polymorphisms: implications for the evolution of doubly uniparental inheritance in bivalves. *Evolution* 62:99–106
- Kulkarni M, Smith HE (2008) E1 Ubiquitin-activating enzyme UBA-1 plays multiple roles throughout *C. elegans* development. *PLoS Genet* 4:e1000131
- Kyriakou E, Kravariti L, Vasilopoulos T, Zouros E, Rodakis GC (2015) A protein binding site in the M mitochondrial genome of *Mytilus galloprovincialis* may be responsible for its paternal transmission. *Gene* 562:83–94
- Kyriakou E, Kravariti L, Zouros E, Rodakis GC (2016) No sex-specific protein-binding site in the VD1 of the F mitochondrial genome of the mussel *Mytilus galloprovincialis*. *Gene Rep* 5:148–150
- Levitan Don R (2000) Sperm velocity and longevity trade off each other and influence fertilization in the sea urchin *Lytechinus variegatus*. *Proc Roy Sc Lond B* 267:531–534
- Lee Y, Kwak H, Shin J, Kim SC, Kim T, Park JK (2019) A mitochondrial genome phylogeny of Mytilidae (Bivalvia: Mytilida). *Mol Phylogenet Evol* 139:106533
- Liu G, Innes D, Thompson RJ (2011) Quantitative analysis of sperm plane circular movement in the blue mussels *Mytilus edulis*, *M. trossulus* and their hybrids. *J Exp Zool A* 315A:280–290
- Lopes-Lima M, Burlakova LE, Karatayev AY, Mehler K, Seddon M, Sousa R (2018) Conservation of freshwater bivalves at the global scale: diversity, threats and research needs. *Hydrobiologia* 810:1–14
- Lopes-Lima M, Froufe E, Ghamizi M, Mock KE, Kebapçı Ü, Klishko O, Kovitvadhi S, Kovitvadhi U, Paulo OS, Pfeiffer JM III, Raley M (2017) Phylogeny of the most species-rich freshwater bivalve family (Bivalvia: Unionida: Unionidae): Defining modern subfamilies and tribes. *Mol Phylogenet Evol* 106:174–191
- Lopes-Lima M, Teixeira A, Froufe E, Lopes A, Varandas S, Sousa R (2014) Biology and conservation of freshwater bivalves: past, present and future perspectives. *Hydrobiologia* 735(1):1–3
- Lubošný M, Przyłucka A, Śmietanka B, Breton S, Burzyński A (2018) Actively transcribed and expressed atp8 gene in *Mytilus edulis* mussels. *PeerJ* 6:e4897
- Lubošný M, Śmietanka B, Przyłucka A, Burzyński A (2020) Highly divergent mitogenomes of *Geukensia demissa* (Bivalvia, Mytilidae) with extreme AT content. *J Zool Syst Evol Res* <https://doi.org/10.1111/jzs.12354>
- Lymbery RA, Kennington WJ, Evans JP (2017) Egg chemoattractants moderate intraspecific sperm competition. *Evol Lett* 1:317–327
- McCartney MA, Auch B, Kono T, Mallez S, Zhang Y, Obille A, Becker A, Abrahante JE, Garbe J, Badalamenti JP, Herman A (2019) The Genome of the Zebra Mussel, *Dreissena polymorpha*: A Resource for Invasive Species Research. *BioRxiv* 1:696732
- McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucl Acids Res* 32:W20–W25
- Milani L (2015) Mitochondrial membrane potential: a trait involved in organelle inheritance? *Biol Lett* 11(10)
- Milani L, Ghiselli F (2015) Mitochondrial activity in gametes and transmission of viable mtDNA. *Biol Direct* 10:22

- Milani L, Ghiselli F, Guerra D, Breton S, Passamonti M (2013) A comparative analysis of mitochondrial ORFans: new clues on their origin and role in species with doubly uniparental inheritance of mitochondria. *Genome Biol Evol* 5:1408–1434
- Milani L, Ghiselli F, Maurizii MG, Nuzhdin SV, Passamonti M (2014) Paternally transmitted mitochondria express a new gene of potential viral origin. *Gen Biol Evol* 6:391–405
- Mitchell A, Guerra D, Stewart D, Breton S (2016) *In silico* analyses of mitochondrial ORFans in freshwater mussels (Bivalvia: Unionoida) provide a framework for future studies of their origin and function. *BMC Genom* 17:597
- Mizi A, Zouros E, Moschonas N, Rodakis GC (2005) The complete maternal and paternal mitochondrial genomes of the Mediterranean mussel *Mytilus galloprovincialis*: implications for the doubly uniparental inheritance mode of mtDNA. *Mol Biol Evol* 22:952–967
- Oliver M, Evans JP (2014) Chemically moderated gamete preferences predict offspring fitness in a broadcast spawning invertebrate. *Proc Biol Sci* 281:20140148
- Ouimet P, Kienzle L, Lubosny M, Burzyński A, Angers A, Breton S (2020) The ORF in the control region of the female-transmitted *Mytilus* mtDNA codes for a protein. *Gene* 725:144161
- Pante E, Poitrimol C, Saunier A, Becquet V, Garcia P (2017) Putative sex-linked heteroplasmy in the tellinid bivalve *Limecola balthica* (Linnaeus, 1758). *J Moll Stud* 83:226–228
- Passamonti M, Ghiselli F (2009) Doubly uniparental inheritance: two mitochondrial genomes, one precious model for organelle DNA inheritance and evolution. *DNA Cell Biol* 28:79–89
- Passamonti M, Ricci A, Milani L, Ghiselli F (2011) Mitochondrial genomes and doubly uniparental inheritance: new insights from *Musculista senhousia* sex-linked mitochondrial DNAs (Bivalvia Mytilidae). *BMC Genom* 12:442
- Passamonti M, Plazzi F (2020) Doubly Uniparental Inheritance and beyond: the contribution of the Manila clam *Ruditapes philippinarum*. *J Zool Syst Evol Res*. <https://doi.org/10.1111/jzs.12371>
- Plazzi F, Passamonti M (2019) Footprints of unconventional mitochondrial inheritance in bivalve phylogeny: signatures of positive selection on clades with doubly uniparental inheritance. *J Zool Syst Evol Res* 57:258–271
- Plazzi F, Puccio G, Passamonti M (2016) Comparative large-scale mitogenomics evidences clade-specific evolutionary trends in mitochondrial DNAs of bivalvia. *Genome Biol Evol* 8:2544–2564
- Plazzi F, Ribani A, Passamonti M (2013) The complete mitochondrial genome of *Solemya velum* (Mollusca: Bivalvia) and its relationships with Conchifera. *BMC Genom* 14:409
- Pozzi A, Plazzi F, Milani L, Ghiselli F, Passamonti M (2017) SmithRNAs: could mitochondria “bend” nuclear regulation? *Mol Biol Evol* 34:1960–1973
- Régnier C, Fontaine B, Bouchet P (2009) Not knowing, not recording, not listing: numerous unnoticed mollusk extinctions. *Cons Biol* 23:1214–1221
- Reis J, Machordom A, Araujo R (2013) Morphological and molecular diversity of Unionidae (Mollusca, Bivalvia) from Portugal. <https://doi.org/10.3989/graellsia.2013.v69.075>
- Riccardi N, Froufe E, Bogan AE, Zieritz A, Teixeira A, Vanetti I, Varandas S, Zaccara S, Nagel KO, Lopes-Lima M (2019) Phylogeny of European Anodontini (Bivalvia: Unionidae) with a re-description of *Anodonta exulcerata*. *Zool J Linn Soc*. <https://doi.org/10.1093/zoolin/zl136>
- Robicheau BM, Breton S, Stewart DT (2017a) Sequence motifs associated with paternal transmission of mitochondrial DNA in the horse mussel, *Modiolus modiolus* (Bivalvia: Mytilidae). *Gene* 605:32–42
- Robicheau BM, Powell AE, Del Bel L, Breton S, Stewart DT (2017b) Evidence for extreme sequence divergence between the male- and female-transmitted mitochondrial genomes in the bivalve mollusc, *Modiolus modiolus* (Mytilidae). *J Zool Syst Evol Res* 55:89–97
- Robicheau BM, Chase EE, Hoeh WR, Harris JL, Stewart DT, Breton S (2018) Evaluating the utility of the female-specific mitochondrial *f-orf* gene for population genetic, phylogeographic and systematic studies in freshwater mussels (Bivalvia: Unionida). *PeerJ* 6:e5007
- Sato M, Sato K (2013) Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. *Biochim Biophys Acta—Mol Cell Res* 1833:1979–1984

- Shea CP, Peterson JT, Wisniewski JM, Johnson NA (2011) Misidentification of freshwater mussel species (Bivalvia: Unionidae): contributing factors, management implications, and potential solutions. *J North Amer Benth Soc* 30:446–458
- Shi J, Hong Y, Sheng J, Peng K, Wang J (2015) *De novo* transcriptome sequencing to identify the sex-determination genes in *Hyriopsis schlegelii*. *Biosci Biotech Biochem* 79:1257–1265
- Śmietanka B, Burzyński A, Wenne R (2010) Comparative genomics of marine mussels (*Mytilus* spp.) gender associated mtDNA: rapidly evolving atp8. *J Mol Evol* 71:385–400
- Śmietanka B, Lubośny M, Przyłucka A, Gérard K, Burzyński A (2018) Mitogenomics of *Perumytilus purpuratus* (Bivalvia: Mytilidae) and its implications for doubly uniparental inheritance of mitochondria. *PeerJ* 6:e5593
- Smith DR (2006) Survey design for detecting rare freshwater mussels. *J North Amer Benth Soc* 25:701–711
- Soroka M, Burzyński A (2017) Hermaphroditic freshwater mussel *Anodonta cygnea* does not have supranumerary open reading frames in the mitogenome. *Mitochondrial DNA Part B* 2:862–864
- Stewart DT, Breton S, Blier PU, Hoeh WR (2009) Masculinization events and doubly uniparental inheritance of mitochondrial DNA: a model for understanding the evolutionary dynamics of gender-associated mtDNA in mussels. *Evolutionary Biology*. Springer, Berlin, Heidelberg, pp 163–173
- Stewart DT, Jha M, Breton S, Hoeh WR, Blier PU (2012) No effect of sperm interactions or egg homogenate on sperm velocity in the blue mussel, *Mytilus edulis* (Bivalvia: Mytilidae). *Can J Zool* 90:1291–1296
- Vallès Y, Boore JL (2006) Lophotrochozoan mitochondrial genomes. *Integr Comp Biol* 46:544–557
- Zbawicka M, Skibinski D, Wenne R (2003) Doubly uniparental transmission of mitochondrial DNA length variants in the mussel *Mytilus trossulus*. *Mar Biol* 142:455–460
- Zieritz A, Geist J, Gum B (2014) Spatio-temporal distribution patterns of three stream-dwelling freshwater mussel species: towards a strategy for representative surveys. *Hydrobiologia* 735:123–136
- Zieritz A, Hoffman JI, Amos W, Aldridge DC (2010) Phenotypic plasticity and genetic isolation-by-distance in the freshwater mussel *Unio pictorum* (Mollusca: Unionoida). *Evol Ecol* 24:923–938
- Zouros E, Freeman KR, Ball AO, Pogson GH (1992) Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel *Mytilus*. *Nature* 359:412–414
- Zouros E (2013) Biparental inheritance through uniparental transmission: the doubly uniparental inheritance (DUI) of mitochondrial DNA. *Evol Biol* 40:1–31
- Zouros E, Rodakis GC (2019) Doubly uniparental inheritance of mtDNA: an unappreciated defiance of a general rule. Cellular and molecular basis of mitochondrial inheritance. Springer, Cham, pp 25–49
- Zouros E (2020) Doubly uniparental inheritance of mitochondrial DNA: Might it be simpler than we thought? *J Zool Syst Evol Res*. <https://doi.org/10.1111/jzs.12364>

# Chapter 13

## The Evolution of the *FLOWERING LOCUS T-Like (FTL)* Genes in the Goosefoot Subfamily *Chenopodioideae*



Helena Štorchová

**Abstract** The assembly of the complete genome of the important crop *Chenopodium quinoa* made possible to identify and analyze the sequences of important regulatory genes. In this review, we focused on the *FLOWERING LOCUS T-like (FTL)* genes—the essential factors controlling flowering in angiosperms. *Chenopodium quinoa* is a tetraploid, which harbors two homeolog copies of many genes including *FTL*s. We recognized seven *FTL* paralogs in *C. quinoa*, each of them existing in two homeolog duplicates. We constructed the phylogenetic tree depicting the relationship of *C. quinoa FTL* genes. We also discussed their evolution in the context of the evolution of *FTL* genes in flowering plants, in particular in the subfamily *Chenopodioideae*.

### 13.1 Introduction

Most crop plants are polyploid (Hilu 1993), which complicates the correct annotation of their genes. The tetraploid *C. quinoa*, which complete genome was published by Jarvis et al. (2017), is not an exception. The current annotation of its *FTL* genes available on GenBank is often incomplete and confounding, which affected for example the recent study of the flowering-related genes in *C. quinoa* (Golicz et al. 2019), which adopted some inaccurate information from GenBank. The orientation among *FTL* paralogs, orthologs, and homeologs in Amaranthaceae was made harder by the confounding names of two essential genes functioning as floral inhibitor and floral activator. They were termed *BvFT1* (inhibitor) and *BvFT2* (activator) by Pin et al. (2010), who revealed their functions in sugar beet. However, the two *FTL* paralogs in Amaranthaceae had been described previously in *Chenopodium rubrum* by Cháb et al. (2008). The floral activator homologous to *BvFT2* was named *CrFTL1*, and the ortholog of *BvFT1* was named *CrFTL2*. Pin et al. (2010) did not refer to the paper by Cháb et al. (2008). The opposite numbering of the two very important genes made

---

H. Štorchová (✉)

Plant Reproduction Laboratory, Institute of Experimental Botany, Czech Academy of Sciences, Rozvojová 263, 16502 Prague, Czech Republic  
e-mail: [storchova@ueb.cas.cz](mailto:storchova@ueb.cas.cz)

the orientation in a very complex set of *FTL* genes in Amaranthaceae even more difficult. To clarify the identity of the *FTL* genes in the very important crop *C. quinoa*, we performed a detailed analysis of its genome, mined all relevant *FTL* sequences, and estimated their phylogenetic relationships. To provide a deeper insight into the evolution of *FTL* genes in Amaranthaceae, I wrote the broader review following the fate of *FTL* genes across seed plants and angiosperms.

## 13.2 The Evolution of *FT*-Like Genes in Seed Plants

The evolution by gene duplication with subsequent gains or modifications of gene function has been recognized for a long time (Ohno 1970). Gene duplication is a fundamental process operating in many gene families of flowering plants including the gene family *FLOWERING LOCUS T/TERMINAL FLOWER1 (FT/TFL1)* comprising important developmental regulators affecting seed germination, flowering time, or plant architecture. The FT/TFL1 proteins are similar in structure to animal phosphatidyl ethanolamine-binding proteins (PEBP). They may be further subdivided into three clades: MFT-like, TFL1-like, and FT-like (Kalgren et al. 2011).

The MFT-like subfamily is supposed to be the ancestral clade. The PEBP genes identified in the moss *Physcomitrella patens* and the lycophyte *Selaginella pallelescens* (Hedman et al. 2009) are more closely related to MFT-like genes than to other subfamilies. However, they are considered to be paralogs rather than direct ancestors of the MFT-like branch of seed plants (Liu et al. 2016). The MFT genes control seed germination in angiosperms (Xi et al. 2010; Yu et al. 2019) and embryo development in gymnosperms (Kalgren et al. 2011). The ancestral function of PEBPs in land plants is not known, but it may be related to the control of plant growth to cope with Earth's gravity after colonizing the land.

The TFL1 gene controls inflorescence architecture through maintaining shoot meristem indeterminacy and acts as floral repressor in *Arabidopsis thaliana* (Wickland and Hanzawa 2015; Perilleux et al. 2019). It also functions as the inhibitor of recurrent flowering in Rosaceae (Wang et al. 2012), or it determines leaf shape in tomato (Lifschitz et al. 2014). Other members of this clade in *A. thaliana* *BROTHER OF FT AND TFL1 (BFT)* and *Arabidopsis thaliana* *CENTRORADIALIS* homolog (*ATC*) suppress flowering under salt stress (Ryu et al. 2011) or under short days (Huang et al. 2012), respectively.

The FT-like subfamily comprises floral activators known as “florigens” (Corbezier et al. 2007), but also genes inhibiting flowering (Pin et al. 2010; Harig et al. 2012; Coelho et al. 2014; Liu et al. 2018) or influencing developmental processes other than flowering (Navarro et al. 2011; Lee et al. 2013). The FT-like genes of gymnosperms control growth rhythms and participate in female and male cone development (Liu et al. 2016). The heterologous expression of spruce FT genes in *A. thaliana* suppresses flowering (Klintonäs et al. 2012), which suggests that only the angiosperm FT genes are capable to activate flowering. However, this experiment

also documents high functional conservancy of *FT* genes, which influence developmental processes in species as distant as spruce and *A. thaliana*, which diverged more than 300 Mya.

The *TFL1*-like and *FT*-like clades arose owing to the gene duplication in the common ancestor of seed plants before gymnosperms split from angiosperms (Liu et al. 2016). Subsequent duplications of *FT/TFL1* genes created additional gene copies in the course of evolution of the two main branches of seed plants. The rapid radiation of angiosperms, which started in the middle Cretaceous period, is tightly associated with whole-genome duplications (WGS) (Wu et al. 2020). Dicots and monocots underwent WGD soon after their divergence— $\gamma$  polyploidization in dicots (Bowers et al. 2003), and  $\tau$  polyploidization in monocots (Jiao et al. 2014). Four phylogenetic branches of *FT-like* gene are present in monocots (Qin et al. 2019), but all dicot *FT* genes belong to a single branch comprising, e.g., *OsHd3a* from rice (Kojima et al. 2002) or *BdFT1* and *BdFT2* genes from *Brachypodium distachyon* (Lv et al. 2014). The retention of *FT* duplicates following the ancient  $\tau$  polyploidization in monocots, but not following the  $\gamma$  polyploidization in dicots, might have been responsible for a higher number of *FT* paralogs identified in monocots (Meng et al. 2011; Zhu et al. 2017) than in dicots.

More recent *FT* gene duplications occurred in particular families or genera of flowering plants and often led to the acquisition of novel functions. Some *FT* copies gained an opposite role of floral inhibitors, e.g., in tobacco (Harig et al. 2012), or in soybean (Liu et al. 2018). Other *FT* paralogs regulate flowering under specific conditions, e.g., *TWIN SISTER OF FT (TSF)* in Brassicaceae, which promotes flowering under short days in response to cytokinins (D'Aloia et al. 2011).

### 13.3 The Diversification of the *FT*-Like Genes in Amaranthaceae

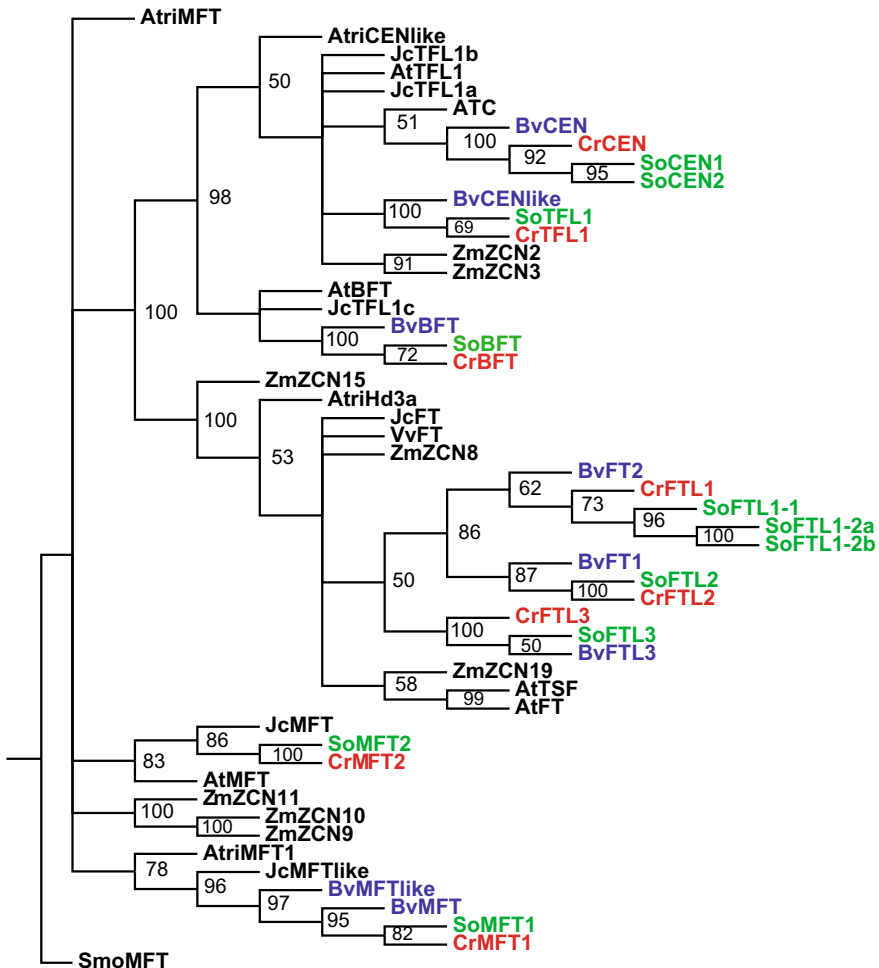
Caryophyllales are the sister of asterids, diverging from a common ancestor early during the evolution of eudicots (Hilu et al. 2003; Soltis et al. 2011; Leebens-Mack et al. 2019). They contain numerous families—e.g., Cactaceae, Aizoaceae, Portulacaceae, Droseraceae, or Nepenthaceae, and some are carnivorous plant families. Most families of Caryophyllales are poorly investigated at genomic level. The family Amaranthaceae represents an exception, because it contains important crops—sugar beet and *Chenopodium quinoa*—with fully sequenced and annotated genomes (Dohm et al. 2014; Jarvis et al. 2017, respectively). The availability of complete genomic sequences makes possible to identify all important gene paralogs including the *FT*-like genes in *C. quinoa*.

The duplication of the *FT* gene occurred at about the origin of Amaranthaceae and led to two phylogenetic clades described first in *Chenopodium rubrum* by Cháb et al. (2008). The *CrFTL1* activated flowering, and the *CrFTL2* gene did not influence flowering despite having been highly expressed (Drabešová et al. 2014). The orthologs



of the two genes were later described in sugar beet, where *BvFT2* (the ortholog of *CrFTL1*) promoted flowering, but *BvFT1* (the ortholog of *CrFTL2*) functioned as floral suppressor (Pin et al. 2010).

Another *FT* paralog named *FTL3* was found in the genome of sugar beet and *C. quinoa*, as well as in genomic DNA of *C. rubrum* (Drabešová et al. 2016) (Fig. 13.1). It diverged much earlier than the *FTL1* and *FTL2* genes. *CrFTL3* was *C. rubrum*, and



**Fig. 13.1** The maximum-likelihood (ML) phylogenetic tree of the FT/TFL1 genes in angiosperms constructed by RAxML v. 8.2.10 (Stamatakis 2014). Bootstrap support of the majority rule consensus tree was calculated from 1000 pseudoreplicates. Species abbreviations: Cr—*C. rubrum*, Bv—*Beta vulgaris*, So—*Spinacia oleracea*, At—*A. thaliana*; Atri—*Amborella trichopoda*; Jc—*Jatropha curcas*, Smo—*Selaginella moellendorffii*, Zm—*Zea mays*. *SmoMFT* was used as outgroup (Drabešová et al. 2016)

only low transcript abundance was detected in seeds and germinating seedlings of this species. It is not known whether *FTL3* occurs in other families beyond Amaranthaceae because of its general absence in transcriptomes and the scarcity of complete genomic sequences in Caryophyllales. The low expression excludes the participation of the *FTL3* gene in the control of flowering, but its function in embryogenesis or germination cannot be excluded considering its expression in seeds (Drabešová et al. 2016).

### 13.4 The *FT-Like* Genes in *Chenopodium Quinoa* Are Numerous

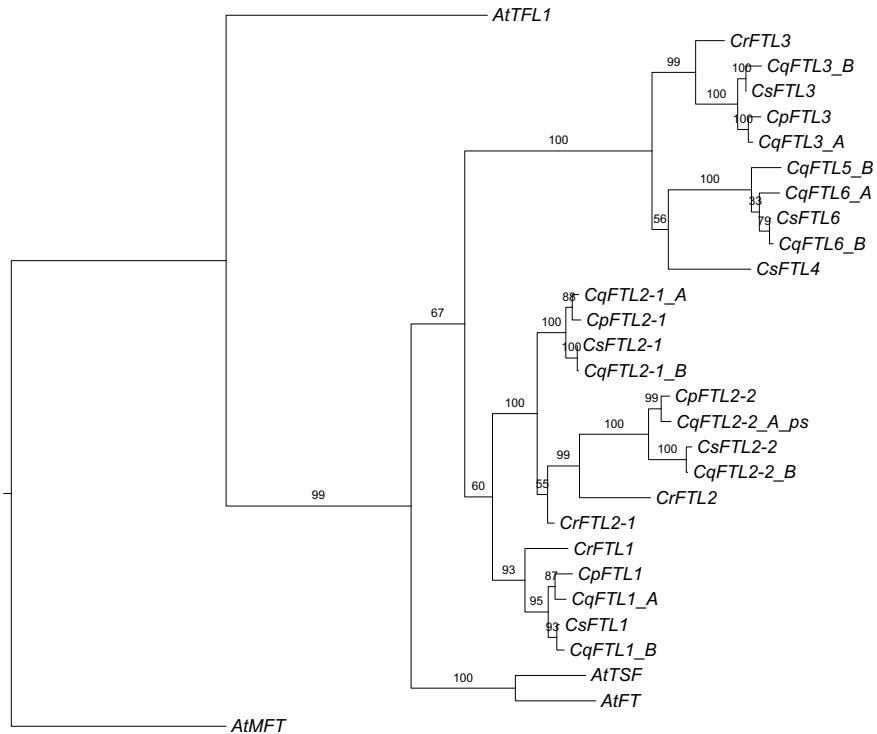
In addition to the ancient *FT* duplicates found in all species of Amaranthaceae so far investigated, more recent *FT* paralogs were observed in some phylogenetic branches of this family. The recent *FTL1* duplications were identified in spinach, but not in sugar beet or *Chenopodium* (Drabešová et al. 2016) (Fig. 13.1), where only a single copy of this floral activator was detected. In contrast, the *FTL2* gene generated two paralogs *FTL2-1* and *FTL2-2* in *Chenopodium* after the divergence of the ancestor of *Chenopodium* species from the ancestor of sugar beet, in which only one *FTL2* gene was found (Pin et al. 2010; Štorchová et al. 2019). The *FTL2-2* copy underwent structural evolution and acquired an additional exon. The expression patterns of the two *FTL2* genes are highly variable. The *FTL2-1* gene of *C. rubrum* was not transcribed at all, whereas the *FTL2-2* gene exhibited high invariant expression (Cháb et al. 2008). The *FTL2-1* and *FTL2-2* genes of *Chenopodium ficifolium*, a close diploid relative of the tetraploid crop *C. quinoa*, showed circadian rhythmicity of transcript levels but differed in the size of amplitude (Štorchová et al. 2019). The changes in gene structure and expression pattern of *FTL2* paralogs suggest the shifts in the function, albeit currently unknown.

The genome of the tetraploid crop *C. quinoa* contains two homeologous copies of each of the *FTL2* genes, four *FTL2* paralogs altogether (Jarvis et al. 2017). The *FTL2-2* copy in the subgenome A derived from the species related to *Chenopodium pallidicaule* is truncated and likely represents a pseudogene, and other three paralogs have complete open reading frames (Table 13.1).

The detailed search for *FTL3* homologs in *C. quinoa* genome revealed several related genes. Their locations in the genome are summarized in Table 13.1, and their phylogenetic relationship is depicted in Fig. 13.2. The *FTL3* orthologs were found in both subgenomes of quinoa, as well as in the diploid species *Chenopodium suecicum* and *C. pallidicaule* related to the donors of subgenome A and subgenome B. They were not truncated and likely capable to encode functional proteins. Figure 13.2 shows similar topology with early branching *FTL3* genes as depicted in Fig. 13.1. However, Fig. 13.1 contains only one copy of *FTL2*, because the *FTL2* duplication was not known previously (Drabešová et al. 2016). The topology of Fig. 13.2 is generally congruent with the species tree of tetraploid *C. quinoa* and its diploid

**Table 13.1** Chromosome positions, genome coordinates, and locus identifications of the *FT*-like genes in *C. quinoa*

Gene	Locus	Chr	Coordinates		Exons	GenBank name	Note
			Start	End			
CqFTL1	LOC110736385	12 A	3,192,504	3,196,018	4	FLOWERING LOCUS T-like	
	LOC110701053	05 B	77,601,501	77,596,889	4	FLOWERING LOCUS T-like	
CqFTL2-1	LOC110718547	15 A	4,933,952	4,930,835	4	FLOWERING LOCUS T-like	
	LOC110739405	17 B	79,266,951	79,270,027	4	FLOWERING LOCUS T-like	
CqFTL2-2	NA	15 A	4,917,910	4,920,973	3		Pseudogene, exon 3 lost
	LOC110739397	17 B	79,273,935	79,277,385	5	HEADING DATE 3A-like	
CqFTL3	LOC110698003	14 A	23,254,495	23,258,984	4	TWIN SISTER of FT-like	
	LOC110713631	06 B	69,532,880	69,534,843	4	TWIN SISTER of FT-like	
CqFTL4	NA	14 A	23,111,694	23,119,151	2		2 exons lost, pseudogene
	LOC110713724	06 B	69,426,761	69,434,601	5	Uncharacterized LOC110713724	
CqFTL5	NA	14 A	22,984,817	22,990,277	2		2 exons lost, pseudogene
	LOC110713625	06 B	69,126,739	69,131,119	4	TWIN SISTER of FT-like	
CqFTL6	LOC110697999	14 A	22,669,634	22,672,839	4	RICE FLOWERING LOCUS T 1-like	
	LOC110713802	06 B	68,781,707	68,786,391	4	TWIN SISTER of FT-like	



**Fig. 13.2** The ML phylogenetic tree of the *FTL*-like genes in the genus *Chenopodium*. *A. thaliana* genes *AtMFT* and *AtFTL1* were used as outgroups. Species abbreviations: At—*A. thaliana*, Cr—*C. rubrum*, Cs—*C. suecicum*, Cp—*C. pallidicaule*, Cq—*C. quinoa*, A—A subgenome of *C. quinoa*, B—B subgenome of *C. quinoa*. Bootstrap support is given above the respective branches

relatives (Štorchová et al. 2015; Mandák et al. 2018), which suggests comparable gene evolutionary rates and no introgression. However, gene *FTL* losses or duplications are common in the *Chenopodium* relatives. For example, spinach or sugar beet does not have two *FTL2* duplicates.

In addition to the *FTL3* orthologs, the sequences clustering with *FTL3*, but forming a different subclade, were found in the quinoa genome. They were named *FTL4*, *FTL5*, and *FTL6* (Table 13.1; Fig. 13.2). Some of those genes lacked one or two exons, or contained frameshift mutations and probably became pseudogenes. The search in the completely sequenced genomes of the diploid relatives of quinoa identified *FTL4* and *FTL6* homologs in *C. suecicum*, but no *FTL4*–*FTL6* homologs in *C. pallidicaule*. Some of the *FTL3*–*FTL6* genes were annotated in the quinoa genome, but their names in GenBank are misleading (*RICE FLOWERING LOCUS T 1-like*, *TWIN SISTER of FT-like*). The assignment of particular genes to the current GenBank accessions is given in Table 13.1.

The evolution of the *FT*-like genes in Chenopodioideae is accompanied by a long history of gene duplications and functional shifts followed by gene losses and pseudogenization. Whereas *FTL1* and *FTL2* homologs participate in the regulation of flowering (Cháb et al. 2008; Pin et al. 2010; Drabešová et al. 2014), the function of the *FTL3* genes is unknown. The *FTL3* clade diversified into several copies, but many of them were subsequently truncated or lost, as documented by the sequences recognized in the genome of *C. quinoa*. The function of *FTL3* and its paralogs is currently unknown, but the expression of *CrFTL3* in seeds and germinating seedlings in *C. rubrum* (Drabešová et al. 2016) suggests their possible function in seed maturation, embryogenesis, or germination. They might have played a role in the adaptation of *C. quinoa* to highly variable environments in its vast geographic distribution area, or in the process of domestication. However, additional experimental data are necessary to verify these hypotheses.

The number of the *FT*-like genes in *C. quinoa* is higher compared with, e.g., sugar beet, which contains only a single *FTL1* copy (*BvFT2*), *FTL2* copy (*BvFT1*) (Pin et al. 2010), and *FTL3* copy (Drabešová et al. 2016). Sugar beet is a diploid, but the high number of *FTLs* in *C. quinoa* cannot be explained only by its tetraploid genome. Sugar beet lacks the *FTL2-1* and *FTL2-2* duplicates, as well as the *FTL4*, 5, and 6 genes.

More research is required to clarify the function of *FT*-like genes in *C. quinoa*, the essential staple crop in Latin America. The knowledge of the control of developmental processes in this species has utmost importance in agriculture.

## 13.5 Summary

The correct annotation of the newly assembled genomic sequences is the basic prerequisite for the clarification of gene function. We described seven homeologous pairs of the *FTL* genes in the genome of *C. quinoa*, and three genes were pseudogenized. The phylogenetic analysis of the *FTL* genes in Chenopodioideae pointed to gene duplications occurring at various times during the evolutionary history of the family Amaranthaceae and consequently is the main cause of the increase of *FTL* gene number. Gene duplications were sometimes followed by gene losses or pseudogenization. The function of three pairs of *FTL* genes was associated with flowering, whereas the function of other genes remains unknown.

## 13.6 Methods: Phylogenetic Analyses

The nucleotide sequences of the *FTL* genes in the genus *Chenopodium* were aligned using MUSCLE with default parameters, as implemented in Geneious 7.1.5. and extensive manual editing guided by virtual amino acid sequences. The novel exon 1 in the *FTL2-2* genes was excluded from the alignment, and the rest of the *Chenopodium*

*FTL* sequences was reliably aligned even with *A. thaliana* genes sequences owing to their high conservancy. The maximum-likelihood (ML) phylogenetic tree (Fig. 13.2) was constructed using RAXML v. 8.2.10 (Stamatakis 2014) with 1000 bootstraps and the GTRGAMMA model for both bootstrapping and tree inference at the CIPRES portal (Miller et al. 2015). The phylogenetic tree of the *FTL* genes in flowering plants (Fig. 13.1) was taken from (Drabešová et al. 2016). It was constructed using the same methods as described for *Chenopodium FTLs* analysis.

**Acknowledgements** The author thanks to Manuela Krüger and to David Gutierrez-Larruscain for the help with sequence analyses. The very helpful comments and advice of the anonymous reviewer are highly appreciated. This work was supported by the grant of the Grant Agency of the Czech Republic 19-01639S and by the Ministry of Education, Youth and Sports of the Czech Republic (MSMT) grant Inter-Action LTAUSA18.

## References

- Bowers JE, Chapman BA, Rong JK, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–438
- Cháb D, Kolář J, Olson MS, Štorchová H (2008) Two *FLOWERING LOCUS T* (*FT*) homologs in *Chenopodium rubrum* differ in expression patterns. *Planta* 228:929–940
- Coelho CP, Minow MA, Chalfun A, Colasanti J (2014) Putative sugarcane *FT/TFL1* genes delay flowering time and alter reproductive architecture in *Arabidopsis*. *Front Plant Sci* 5:221
- Corbesier L, Vincent C, Jang SH, Fornara F, Fan QZ, Searle I, Giakountis A, Farrona S, Gissot L, Turnbull C, Coupland G (2007) FT protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science* 316:1030–1033
- D’Aloia M, Bonhomme D, Bouche F, Tamseddak K, Ormenese S et al (2011) Cytokinin promotes flowering of *Arabidopsis* via transcriptional activation of the *FT* paralogue *TSF*. *Plant J* 65:972–979
- Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutierrez S, Zakrzewski F et al (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505:546–549
- Drabešová J, Cháb D, Kolář J, Haškovcová K, Štorchová H (2014) A dark-light transition triggers expression of the floral promoter *CrFTL1* and downregulates *CONSTANS*-like genes in a short-day plant *Chenopodium rubrum*. *J Exp Bot* 65:2137–2146
- Drabešová J, Černá L, Mašterová H, Koloušková P, Potocký M, Štorchová H (2016) The evolution of the *FT/TFL1* genes in Amaranthaceae and their expression patterns in the course of vegetative growth and flowering in *Chenopodium rubrum*. *G3-Genes Genomes Genet* 6:3066–3076
- Golicz AA, Steinfort U, Arya H, Singh MB, Bhalla PL (2019) Analysis of the quinoa genome reveals conservation and divergence of the flowering pathways. *Funct Integr Genomics* 20:245–258
- Harig L, Beinecke FA, Oltmanns J, Muth J, Müller O, Rüping B, Twyman RM, Fischer R, Prüfer D, Noll GA (2012) Proteins from the *FLOWERING LOCUS T*-like subclade of the PEBP family act antagonistically to regulate floral initiation in tobacco. *Plant J* 72:908–921
- Hedman H, Källman T, Lagercrantz U (2009) Early evolution of the *MFT*-like gene family in plants. *Plant Mol Biol* 70:359–369
- Hilu KW (1993) Polyploidy and the evolution of domesticated plants. *Am J Bot* 80:1494–1499
- Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS et al (2003) Angiosperm phylogeny based on *matK* sequence information. *Am J Bot* 90:1758–1776
- Huang NC, Jane WN, Chen J, Yu TS (2012) *Arabidopsis thaliana* *CENTRORADIALIS* homologue (*ATC*) acts systemically to inhibit floral initiation in *Arabidopsis*. *Plant J* 72:175–184

- Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel S, Li B et al (2017) The genome of *Chenopodium quinoa*. *Nature* 542:307–312
- Jiao YN, Li JP, Tang HB, Paterson AH (2014) Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 26:2792–2802
- Kalgren A, Gyllestrand N, Källman T, Sundström JF, Moore D, Lascoux M, Lagercrantz U (2011) Evolution of the *PEBP* gene family in plants: functional diversification in seed plant evolution. *Plant Physiol* 156:1967–1977
- Klintonäs M, Pin PA, Benlloch R, Ingvarsson PK, Nilsson O (2012) Analysis of conifer *FLOWERING LOCUS T/TERMINAL FLOWER1*-like genes provides evidence for dramatic biochemical evolution in the angiosperm *FT* lineage. *New Phytol* 196:1260–1273
- Kojima S, Takahashi Y, Kobayashi Y, Monna L, Sasaki T, Araki T, Yano M (2002) *Hd3a*, a rice ortholog of the *Arabidopsis FT* gene, promotes transition to flowering downstream of *Hd1* under short day conditions. *Plant Cell Physiol* 43:1096–1105
- Lee R, Baldwin S, Kenel F, McCallum J, Macknight R (2013) *FLOWERING LOCUS T* genes control onion bulb formation and flowering. *Nat Commun* 4:2884
- Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA et al (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679–685
- Lifshitz E, Ayre BG, Eshed Y (2014) Florigen and anti-florigen—a systemic mechanism for coordinating growth and termination in flowering plants. *Front Plant Sci* 5:465
- Liu YY, Yang KZ, Wei XX, Wang XQ (2016) Revisiting the phosphatidylethanolamine-binding protein (PEBP) gene family reveals cryptic *FLOWERING LOCUS T* gene homologs in gymnosperms and sheds new light on functional evolution. *New Phytol* 212:730–744
- Liu W, Jiang B, Ma L et al (2018) Functional diversification of *Flowering Locus T* homologs in soybean: *GmFT1a* and *GmFT2a/5a* have opposite roles in controlling flowering and maturation. *New Phytol* 217:1335–1345
- Lv B, Nitcher R, Han XL, Wang SY, Ni F, Li K, Pearce S, Wu JJ, Dubcovsky J, Fu DL (2014) Characterization of *FLOWERING LOCUS T1 (FT1)* gene in *Brachypodium* and wheat. *PLoS ONE* 9:e94171
- Mandák B, Krak K, Vít P, Lomonosova MN, Belyayev A, Habibi F, Wang L, Douda J, Štorchová H (2018) Hybridization and polyploidization within the *Chenopodium album* aggregate analysed by means of cytological and molecular markers. *Mol Phylogenet Evol* 129:189–201
- Meng X, Muszynski MG, Danilevskaya ON (2011) The *FT*-like *ZCN8* gene functions as a floral activator and is involved in photoperiod sensitivity in maize. *Plant Cell* 23:942–960
- Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, Passarotti M, Kaufman S, O’Leary, MA (2015) A RESTful API for access to phylogenetic tools via the CIPRES science gateway. *Evol Bioinform* 11:43–48
- Navarro C, Abelenda JA, Cruz-Oró E, Cuéllar CA, Tamaki S, Silva J, Shimamoto K, Prat S (2011) Control of flowering and storage organ formation in potato by *FLOWERING LOCUS T*. *Nature* 478:119–122
- Ohno S (1970) Evolution by gene duplication. Springer, Heidelberg
- Perilleux C, Bouche F, Randoux M, Orman-Ligeza B (2019) Turning meristems into fortresses. *Trends Plant Sci* 24:431–442
- Pin PA, Benlloch R, Bonnet D, Wremerth-Weich E, Kraft T, Gielen JLL, Nilsson O (2010) An antagonistic pair of *FT* homologs mediates the control of flowering time in sugar beet. *Science* 330:1397–1400
- Qin ZR, Bai YX, Muhammad S, Wu X, Deng PC, Wu JJ, An HL, Wu L (2019) Divergent roles of *FT*-like 9 in flowering transition under different day lengths in *Brachypodium distachyon*. *Nature Commun* 10:812
- Ryu JY, Park CM, Seo PJ (2011) The floral repressor *BROTHER OF FT AND TFL1 (BFT)* modulates flowering initiation under high salinity in *Arabidopsis*. *Mol Cells* 32:295–303
- Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC et al (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot* 98:704–730

- Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313
- Štorchová H, Drabešová J, Cháb D, Kolář J, Jellen EN (2015) The introns in *FLOWERING LOCUS T-LIKE (FTL)* genes are useful markers for tracking paternity in tetraploid *Chenopodium quinoa* Willd. *Genet Resour Crop Evol* 62:913–925
- Štorchová H, Hubáčková H, Abeyawardana OAJ, Walterová J, Vondráková Z, Eliášová K, Mandák B (2019) *Chenopodium ficifolium* flowers under long days without upregulation of *FLOWERING LOCUS T (FT)* homologs. *Planta* 250:2111–2125
- Wang LN, Liu YF, Zhang YM, Fang RX, Liu QL (2012) The expression level of *Rosa Terminal Flower 1 (RTFL1)* is related with recurrent flowering in roses. *Mol Biol Rep* 39:3737–3746
- Wickland DP, Hanzawa Y (2015) The *FLOWERING LOCUS T/TERMINAL FLOWER1* gene family: functional evolution and molecular mechanisms. *Mol Plant* 8:983–997
- Wu SD, Han BC, Jiao YN (2020) Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol Plant* 13:59–71
- Xi W, Liu C, Hou X, Yu H (2010) *MOTHER OF FT AND TFL1* regulates seed germination through a negative feedback loop modulating ABA signaling in *Arabidopsis*. *Plant Cell* 22:1733–1748
- Yu XL, Liu H, Sang N, Li YF, Zhang TT, Sun J, Huang XZ (2019) Identification of cotton *MOTHER OF FT AND TFL1* homologs, *GhMFT1* and *GhMFT2*, involved in seed germination. *PLoS ONE* 14:e0215771
- Zhu YJ, Fan YY, Wang K, Huang DR, Liu WZ, Ying JZ, Zhuang JY (2017) Rice *Flowering Locus T1* plays an important role in heading date influencing yield traits in rice. *Sci Rep* 7:4918



# Chapter 14

## DDE Transposon as Public Goods



Louis Tsakou-Ngouafo, Célia Vicari, Laura Helou, Vivek Keshri, Sabyasachi Das, Yves Bigot, and Pierre Pontarotti

**Abstract** DDE Transposons have been recruited at least four times as site-specific recombination activating gene allowing programmed DNA elimination in eukaryotes. The described cases are RAG in jawed vertebrates, Kat 1 and Alpha 3 in the *Kluyveromyces lactis* yeast and Piggymac/TPB1 TPB2 and TPB6 in ciliates. The domesticated RAG is the most known case. It constitutes the enzymatic core of the Jawed vertebrates V(D)J recombination machinery. It directs random assembly and joining of gene segments during the development of B and T cells helping in the generation of the enormous gene diversity encoding antibodies or T cell receptors. It was shown in the case of RAG that the shift from DDE transposon to site-specific recombination activating gene is an evolutionary phenomenon that did not require dramatic changes. This explains why the co-option of DDE transposon as site-specific recombination activating gene can occur in a convergent manner. As numerous genes coding for DDE transposases are widespread through numerous members of the life tree, it is expected that several of them might correspond to domesticated transposons involved in programmed DNA elimination and maybe in the generation of receptor diversity. The domestication of DDE transposon could have been and still be of an extreme importance for organisms' evolution.

---

L. Tsakou-Ngouafo · C. Vicari · V. Keshri · P. Pontarotti (✉)  
Aix Marseille Univ IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille, France  
e-mail: [Pierre.pontarotti@univ-amu.fr](mailto:Pierre.pontarotti@univ-amu.fr)

L. Helou · Y. Bigot  
UMR INRAE 0085, CNRS 7247, PRC, Centre INRAE Val de Loire, 37380 Nouzilly, France

S. Das  
Emory Vaccine Center and Department of Pathology and Laboratory Medicine, Emory University,  
1462 Clifton Road North-East, Atlanta, GA 30322, USA

P. Pontarotti  
SNC5039 CNRS, Paris, France

## 14.1 Preamble

Eukaryotes and prokaryotes do not have the same DNA content in their genome over stages of their cell differentiation. Indeed genomes are modified during development or cell differentiation. This phenomenon is called developmentally regulated genome rearrangement (DRGR) (Zufall et al. 2005). DRGRs start by a DNA breaks either due to a specific endonuclease activity or specific chromatin structure that is related to a mechanism inducing a DNA recombination event.

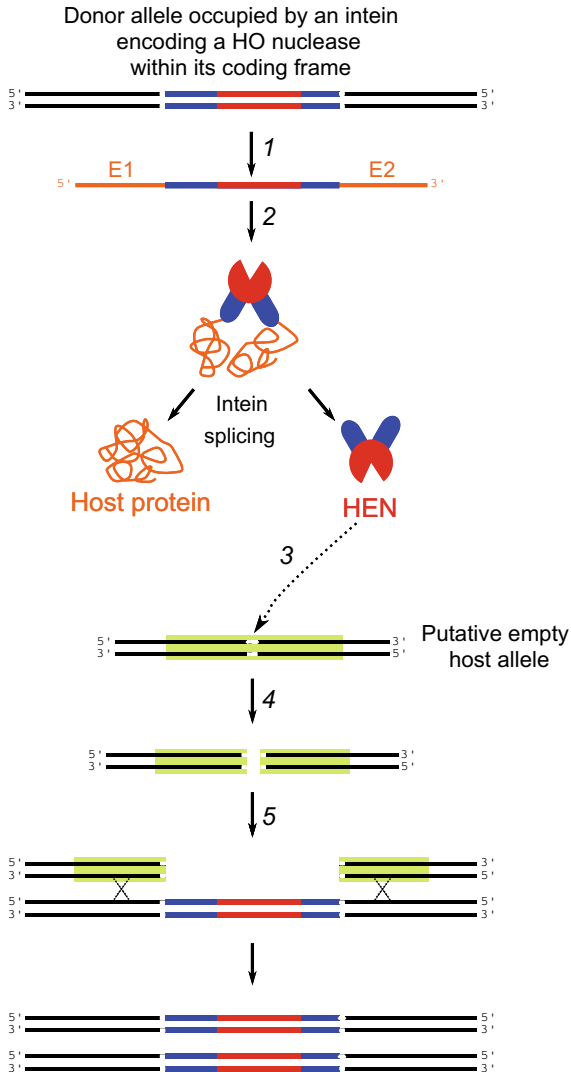
## 14.2 Developmentally Regulated Genome Rearrangement via a Specific Endonuclease

All the reported cases so far correspond to the domestication of selfish DNA that displayed at the minimum an endonuclease activity derived from a DDE transposase, a homing endonuclease, or a prophage recombinase.

**The homothallic switching endonuclease (HO) is a domesticated homing endonuclease** (Fig. 14.1). The HO endonucleases are mobile genetic elements that correspond to DNA fragments encoding these proteins over all their length, from their first 5' to the last 3' nucleotide. These DNA fragments use the homing endonuclease they encode to mediate their mobility (Keeling and Roger 1995; Koufopanou and Burt 2005). Homing endonucleases specifically recognize and cleave a specific oligonucleotide motif that is generally located within a very conserved motif of some house-keeping protein-coding sequence. After DNA cleavage, a DNA fragment coding an HO endonucleases specifically inserts in frame with the parasitized gene in the middle of its own recognition sequences (Fig. 14.1). In a diploid cell that is heterozygous for a homing endonuclease, the gene lacking the parasitic element becomes cleaved at the recognition site, and the broken chromosome is invaded by the parasitic DNA fragment by homologous recombination using the homing endonuclease-containing gene as a template. The mating type in budding yeasts is derived from the domestication of one of these parasitic DNA fragments. The mating types in budding yeasts are encoded by the mating-type (MAT) loci *MAT*<sup>a</sup> and *MAT*<sup>α</sup>. Some yeasts have the ability to change their mating type (mating-type switching) without going through mating or meiosis." In *Saccharomyces cerevisiae*, switching is initiated when the homothallic switching (HO) endonuclease induces a DNA double-strand break (DSB) in the *MAT* locus. Next, the replacement is completed through a gene conversion, in which transcriptionally silent copies of *MAT* genes, known as "hidden MAT left, *HML*<sup>α</sup>" and "hidden MAT right, *HMR*<sup>a</sup>," are copied into the expressed *MAT* locus.

**Prophage excision involved in the specific rearrangement.** (Feiner et al. 2015, Fig. 14.2)

During the differentiation process in prokaryote, prophages and derived prophages are excised from the genomic region using a prophage recombinase in order to

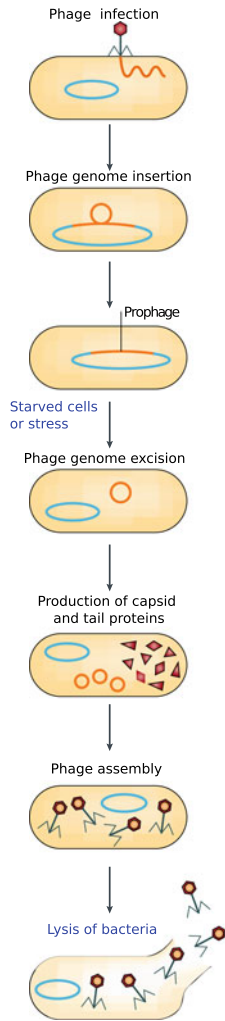


**Fig. 14.1** Organization, expression, splicing, and mobility from one allele occupied by an intein to an unoccupied allele. The main steps of the intein mobility and homing are numbered. Step 1 is the RNA transcription. Black lines indicate DNA strands of the occupied and unoccupied alleles by an intein-coding DNA fragment, and cyan blue and light green bars indicate the coding-frame of the DNA region transcribed for each allele kind. Blue and red lines indicated regions coding for the intein (blue) and the homing endonuclease (HEN) moieties (red) in the DNA and the corresponding RNA transcripts. Step 2 is translation into protein. “pacmans” HEN moieties are shown in red and blue ellipses are N- and C-terminal intein moieties. Step 3 is site-specific recognition of an unoccupied allele by HEN. Blank spaces indicated the insertion site specifically cleaved by the HEN in both alleles. Step 4 is specific cleavage by the HEN. Finally, steps 5 and 6 depicted the invasion of a DNA fragment encoding intein from an occupied allele toward an unoccupied one by DNA strand invasion at DNA replication. This figure was constructed from derived information and graphic elements from Piègu et al. (2015)

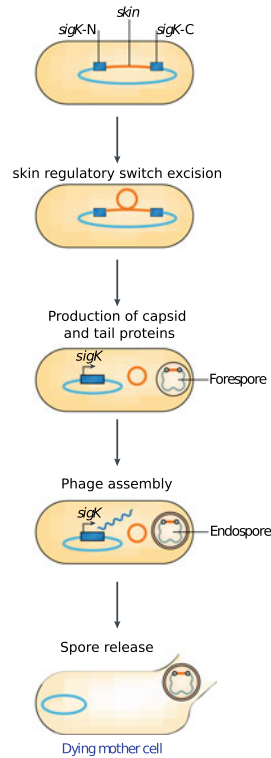
combine ORFs in frame. This process is involved in at least three phenomena: sporulation (*Bacillus* and *Clostridium*), heterocyst differentiation (Cyanobacteria), and monocytogenes phagosomal escape (*Listeria*).

In the case of the “well analyzed” *Bacillus subtilis*, gene rearrangement occurs in the sigK gene that is disrupted into two segments by the insertion of SKIN (sigK intervening element) which is a remnant of the ancestral prophage (Abe et al. 2017a). During sporulation, SKIN is excised from the chromosome to combine the ORFs in

**a. Lysogenic phage cycle**



**b. Spore development in *Bacillus subtilis***



◀**Fig. 14.2** Parallel between the lysogenic cycle of a phage (a) and the bacterial process involving a phage-derived mechanism activating sporulation (b). **a** Temperate phages enter a cycle in which their phage genome is first going to be integrated into the bacterial chromosome to become a prophage that is going to persist in what is considered a phage latent or dormant state in which bacterial cells will be viable and will not produce of phage particles. Prophages are replicated together with the bacterial host chromosome during host-cell replication and switch into lytic production upon exposure to stress. Chronologically, the entry in the lytic phase starts with the excision of the phage genome, its gene transcription then the production of phage capsid and tail proteins that are thereafter assembled in phage particles. The last step is the lysis of the bacteria that releases phages and kills the bacteria. **b** Regulation of mother cell-specific genes during sporulation in *Bacillus subtilis*. A phage regulatory switch (phageRS), named *skin*, is inserted within the *sigK* gene. *SigK* encodes a protein,  $\sigma_K$ , that regulates the expression of late-stage sporulation genes in the mother cell. *skin* excises itself at the initiation of the sporulation process, leaving an intact *sigK* gene able to encode a functional  $\sigma_K$  protein. Post excision,  $\sigma_K$  expression activates the mother cell's late-stage sporulation genes. Following excision, the excised *skin* element is eventually lost in the mother cell, which dies late during sporulation. By contrast, the forespore, which did not undergo element excision, gives rise to an endospore that still encodes the *skin* element within its *sigK* gene. This figure was constructed from derived information and graphic elements from

frame. In addition to *sigK*, many other examples of sporulation-specific gene rearrangement occur, suggesting that this phenomenon is widespread and common in spore-forming bacteria (the intervening sequence can correspond to only a recombinase of prophage origin. The recombinase has two functions: it catalyzes the DNA cleavage at the recombination site and join the DNA molecule ends of the restored protein-coding genes (Abe et al. 2017b) (Fig. 14.2)

Heterocysts in cyanobacteria are specialized cells with a role in nitrogen fixation. They provide nitrogen to vegetative cells. Differentiation in a heterocyst is due the excision of fragments that interrupt three different loci: *nitD*, *fdxN*, and *hupL* that encodes a dinitrogenase alpha subunit, a ferredoxin and an uptake hydrogenase large subunit, respectively. The excised fragments have prophage origins and include at least the recombinase gene (Hilton et al. 2016).

Finally, to promote phagosomal escape, *Listeria monocytogenes* need to excise a temperate prophage integrated into the *comK* locus during bacterial phagocytosis in order to activate the *comK* expression (Feiner et al. 2015)

## 14.3 Programmed Break Specified by Chromatin Signals

### 14.3.1 Mating-Type Switching in the Fission Yeast *S. pombe*

As in the case of *S. cerevisiae*, the yeast *Schizosaccharomyces pombe* is able of mating-type switching. This mechanism occurred in an independent manner, the two species that belong to two highly divergent clades of yeast. The *S. pombe* genome contains one active (*mat1*) and two silenced (*mat2* and *mat3*) mating-type loci that share similarities with the *S. cerevisiae* MAT proteins. However, the mechanism

of switching is different from that involved in *S. cerevisiae*. Instead of cleavage mediated by an HO endonuclease (Fig. 14.2), a fragile chromosomal site consisting of an epigenetic mark at the *mat1* locus in switching-competent cells leads to a dsDNA break during replication (Klar et al. 2014).

### **14.3.2 Immunoglobulin Switching**

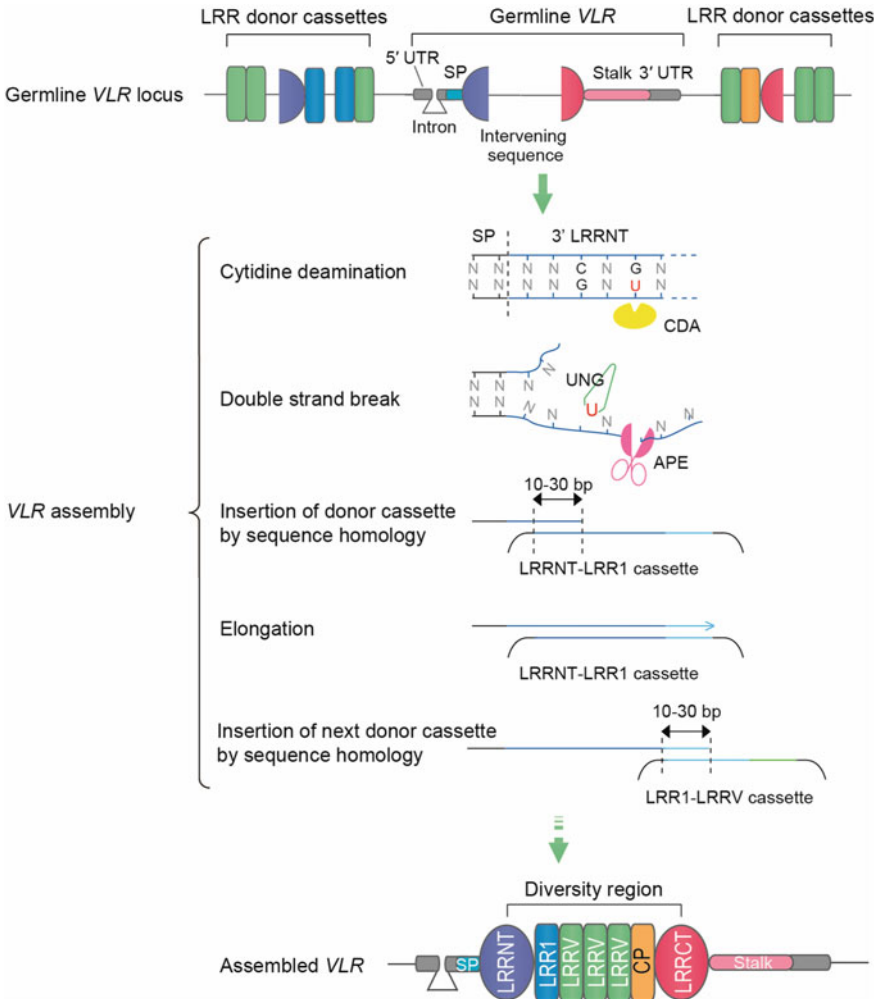
Immunoglobulin isotype class switching is a mechanism that changes a B cell's production of immunoglobulin from one type to another. It involves an intrachromosomal deletional rearrangement that focuses on the region of 1–10 kb of repetitive switch (S) located upstream of each heavy chain isotype gene. Class switch recombination is induced by a break done by cytidine deaminase AID (AID-APOBEC family) on accessible chromatin which is expressed during the B cell development (Yu and Lieber 2019).

### **14.3.3 The variable lymphocyte receptors (VLRs) system (Fig. 14.3)**

In agnathes (cyclostomes), the VLR system involved in adaptive immunity includes a germline VLR that do not code for functional protein but instead encodes the only portion of the amino and carboxyl termini of the mature VLRLs. The sequences encoding those portions are separated by non-coding intervening regions. In lymphocytes, the germline VLRLs are assembled by somatic DNA rearrangement into a mature VLR that encodes the functional receptor via the insertion of LRR cassette that flank the germline VLR. The germline VLR is broken by the AID-APOBEC enzyme at the intervening sequences between the C-terminal and N-terminal portions of the VLR genes where the chromatin is accessible. Then gene conversion starts thanks to sequence identity between the intervening sequence and sequences surrounding the LRR cassettes (Boehm et al. 2018).

### **14.3.4 Immunoglobulin Diversity Driven by Gene Conversion in Birds and Some Mammals**

In birds and some mammals, only one pair of functional V and J segments is found for both the Ig light and heavy chain loci. Therefore, the diversity generated by V(D)J recombination is limited. However, several pseudo-V coding segments are found upstream the functional V segment in genes coding light and heavy chains. These pseudo-V segments are used as a template for gene conversion to diversify the



**Fig. 14.3** Variable lymphocyte receptors (VLRs) system. Schematic diagram of a germline *VLR* gene and the postulated gene conversion-like (copy choice) mechanism for *VLR* assembly. The germline *VLR* gene is incomplete and contains invariant regions encoding for 5'-end of the N-terminal LRR (5'LRRNT) and 3'-end of the C-terminal LRR (3'LRRCT) and stalk. Hundreds of different LRR cassettes are located upstream and downstream of a pre-assembled germline *VLR* gene. The non-coding intervening sequence between 5'LRRNT and 3'LRRCT is replaced by the donor LRR cassettes that are sequentially copied either from 5' to 3' or 3' to 5' direction. At the beginning of the gene assembly process, a cytosine deaminase (CDA) converts cytosine (C) to uracil (U) in the germline *VLR* gene. The uracil is then removed by uracil-DNA glycosylase (UNG), leaving an apurinic (AP) site. The AP site activates nicking activity of AP endonuclease (APE) which leads to a DNA double-strand break. To repair this break, homologous recombination starts based on the sequence homology of 10-30 bp between the donor and acceptor LRR cassettes. This process is repeated along with deletion of the intervening sequence until the completion of a mature *VLR* gene. The diversity region of a mature *VLR* is composed of a 3'LRRNT, LRR1, multiple LRRVs, connecting peptide (CP), and a 5'LRRCT

single functional V segment. To launch the diversification process, a break done by the Activation-induced deaminase (AID) within the single functional V segment is required. (see for example Arakawa et al. 2002).

Other DRGRs have been described in the literature (Wang and Davis 2014), and they likely correspond to an infinitesimal number of cases that occur in nature. At the molecular level, as will be seen in the next section beside the prophage and among other domesticated selfish elements, domesticated DDE transposons seem a solution often retained by evolution to achieve programmed DNA elimination.

## 14.4 DDE Transposon and Domesticated DDE Transposon

DNA transposition is the process by which a discrete segment of DNA is either moved or copied into a new genomic location. Several distinct types of enzymes catalyze DNA transposition, one of the most abundant kinds are the DDE transposases thus named for conserved essential acidic residues located at the active catalytic site. The DDE transposase coding gene in the transposon is flanked by two Terminal Inverted Repeats (TIRs). To achieve transposition, the transposase recognizes these TIRs to perform the excision of the transposon which is after this or in a concerted manner with excision is re-inserted into a new genomic location. Upon insertion, the target site DNA is duplicated, resulting in Target Site Duplications (TSDs).

Several DDE transposons have been domesticated by their hosts retaining their DNA binding capacity (see for review (Sinzelle et al. 2009; Jangam et al. 2017)) and in some cases, the transposase and their related TIR as site-specific recombination activating gene. The few well-studied examples will be presented in the next sections.

### 14.4.1 *Shift from a DDE Transposon to a Mechanism of Programmed DNA Elimination*

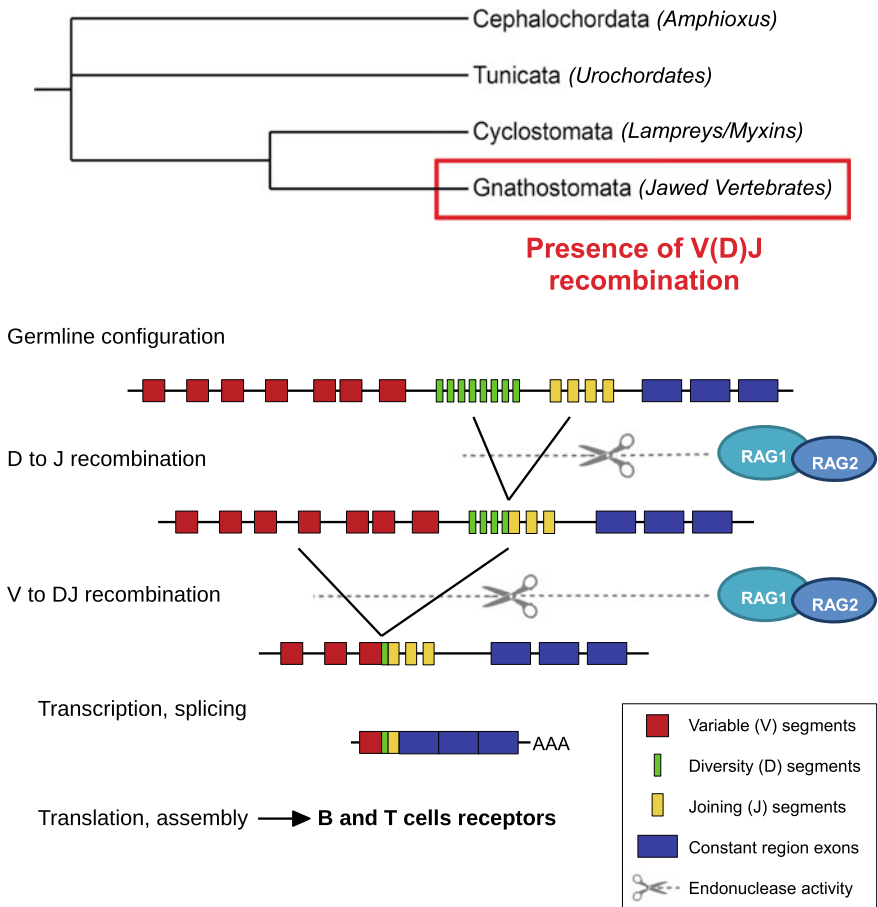
DDE transposons have been recruited as site-specific recombination activating gene at least four times as programmed DNA elimination system, mediating biological differentiation processes: RAG (RAG1-RAG2) in jawed vertebrates, *Kat 1* and *Alpha 3* in *K. lactis* yeast and *PiggyBac* (TPB1 TPB2 and TPB6) in the ciliate *Tetrahymena*. The first described and best known example is the V(D)J RAG.

**The Recombination Activating Gene (RAG) Paradigm** Jawed vertebrates has a specific adaptive immune system based on lymphocytes that express highly diverse, clonally distributed antigen receptors encoded by genes that are non-functional in the germline and assembled by recombination during lymphocyte development (Teng and Schatz 2015). This assembly reaction, known as V(D)J recombination, operates on arrays of V, D, J polypeptide-coding segments of immunoglobulin and T-cell receptor loci. V(D)J recombination is initiated early in lymphocyte development



by a site-specific endonuclease: RAG1 and 2. The RAGs cleave at a conserved recombination signal (RSS) that flanks each V D and J segments. The finding that the RAGs have transposase activity, supports a model in which co-option of the components of a transposon played a critical role in the evolution of the jawed vertebrate adaptive immune system (see for review Flajnik 2016) (Fig. 14.4).

The hypothesis was that the RAG proteins derive from transposase genes of RAG transposons while the split antigen receptor genes derive from the insertion of the terminal inverted repeats (TIRs) of this transposon into a Ig-like receptor gene exons with the inserted TIRs becoming the RSSs. The presence of a RAG transposase core and TIRs in non-vertebrates (Kapitonov and Jurka 2005) was consistent with this model. The discovery of a complete RAG transposon in amphioxus (*Branchiostoma*



**Fig. 14.4** RAG implication in the V(D)J recombination mechanism in jawed vertebrates. The chordates consensus tree shows the position of jawed vertebrates among chordates. The V(D)J mechanism was adapted from Janeway et al. (2001)

*belcheri*) (Huang et al. 2016) strengthens this hypothesis. The discovery of an active RAG transposon in the hemichordata *Ptychodera flava* and several fossilized transposons in several Deuterostomia (Morales Poole et al. 2017) indicate that the RAG transposon was present in the deuterostomia ancestor and remained active to date in several lineages of this clade, while it was co-opted as part of V(D)J recombinase in jawed vertebrates about five hundred millions years ago. The RAG transposon includes TIRs that are related in sequence to the RSS heptamer and the transcribed open reading frame encodes RAG1-like and RAG-2 like proteins with a biochemical activity similar to those of the RAGs, including DNA cleavage via a nick-hairpin mechanism (Huang et al. 2016) (Fig. 14.5).

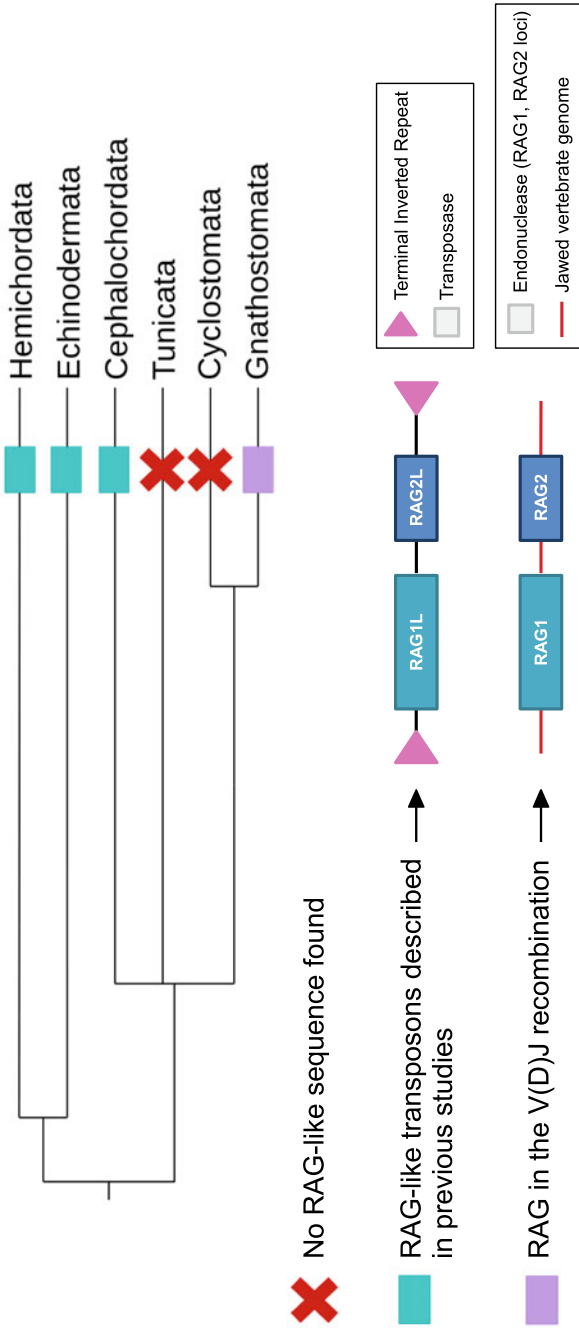
After DNA cleavage by the jawed vertebrate RAG or *Branchiostoma belcheri* RAG transposon, there are two predominant fates for the excised fragment flanked by RSSs or TIRs: joining of the ends to form a signal joint or integration into a new locus. The jawed vertebrate recombinase RAG strongly favors the joining of the ends. Indeed, RAG in the jawed vertebrates actively directs cleaved signals and coding ends into the NHEJ DNA repair pathway for joining coding segments in frame. In contrast, *Branchiostoma belcheri* RAG transposon appears to strongly favor integration after excision, nevertheless, it allows some TIR-TIR joints to form (Huang et al. 2016; Zhang et al. 2019). It is possible that the transposase partially prevents the interaction between the TIR and the NHEJ repair pathway and that the jawed vertebrate RAG lost this property. Zhang et al. (2019) started to uncover the mechanism beyond the domestication and evidenced important amino acid positions involved in transposition/or suppressing transposition (Fig. 14.7).

The mechanism of co-option can be described as follows: (1) insertion of an active transposon into a given genomic locus, (2) the translocation of the transposase gene; in some cases, the native TIRs remain within the genomic locus. (3) The TIR like sequence is recognized by the domesticated transposase (Known under the name of RSSs in the case of RAG) and the TIRs together with the internal sequence are excised (4) the excised sequence lost the ability to insert another genomic region and the two flanking ends are joined by the non-homologous end-joining (NHEJ) DNA repair pathway, to form a coding joint (CJ) (Fig. 14.6).

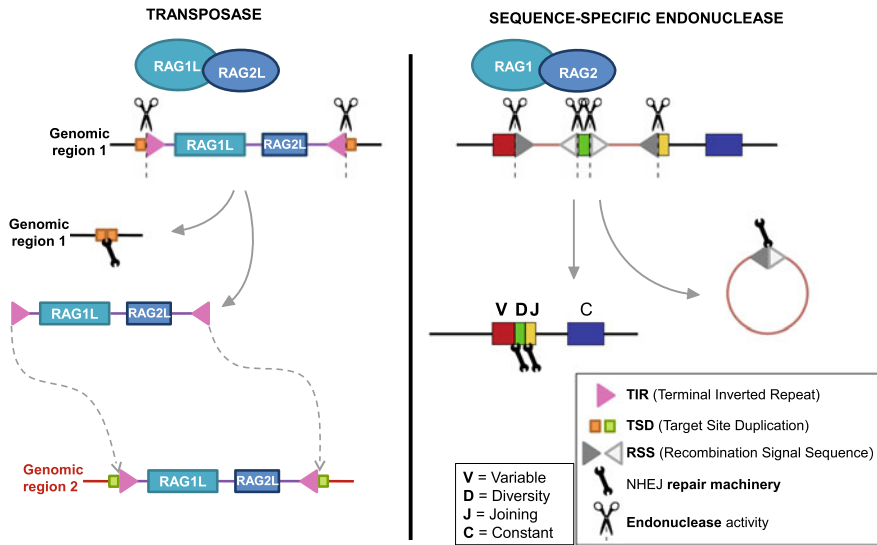
In conclusion, biochemical functions of the DDE transposon and the sequence-specific recombination activating system are very similar, hence the biochemical shift from a transposon to a sequence-specific recombination activating system corresponds an easy evolutionary step requiring a very limited number of events to pass form a “wild” state to a domesticated one (Fig. 14.3). This is also supported by the fact that many other DDE transposons have been co-opted as recombination activating site-specific endonuclease as described in the next paragraphs.

**Piggymac/TPB2/TPB1/TPB6** in ciliates (Baudry et al. 2009; Cheng et al. 2016; Nowacki et al 2009) (Fig. 14.7).

Ciliates are unicellular organisms able to perform DNA rearrangements during development in order to differentiate a somatic macronucleus that is metabolically active from their transcriptionally silent germinal micronucleus. After duplication of the germinal genome, a large proportion of their germinal micronuclear genome is eliminated to differentiate a somatic macronuclear genome through the loss and in



**Fig. 14.5** Taxonomic distribution of the RAG and RAG-like sequences known in Deuterostomes. This is a deuterostomes consensus tree with the clades presenting RAG in the V(D)J recombination in the adaptive immune system, the clades presenting RAG-like sequences and the clades presenting no RAG-like sequences until today, described in Morales Poole et al. (2017)

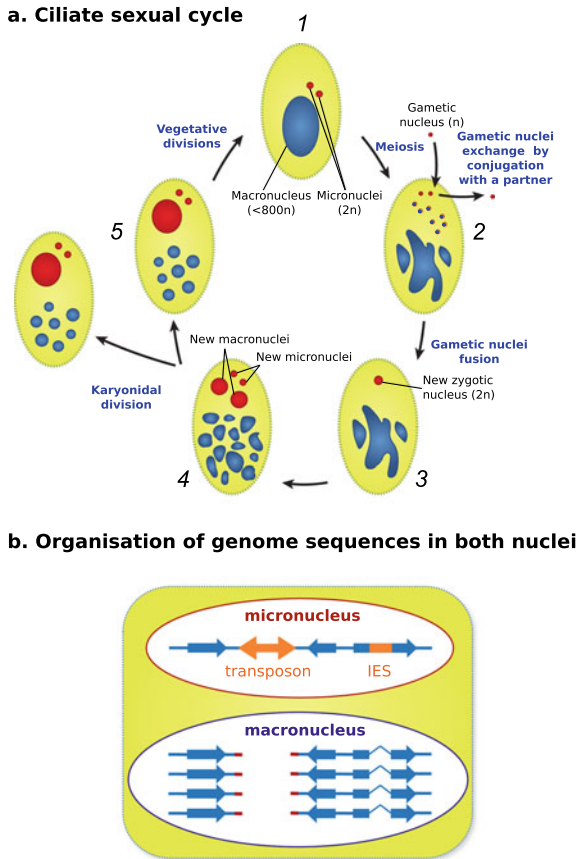


**Fig. 14.6** Simple biochemical switch between the RAG-like transposase activity and the RAG endonuclease/recombinase activity in the V(D)J recombination of jawed vertebrates. Adapted from Huang et al. (2016)

particular through the excisions of thousands of internal eliminated sequences. This is done by domesticated DDE transposon. Most of the eliminated sequences in the paramecium and tetrahymena are named, respectively, by Piggymac and TPB2 and are derived from transposons belonging to the piggybac family (Fig. 14.7). In contrast to the domesticated *piggyBac* transposase in the macronuclear genomes of oligohymenophorea ciliates, *Oxytricha* bears thousands of active transposase genes within the *Tc1/mariner* superfamily.

In the case of the Tetrahymena, the domesticated transposons TPB1 and TPB6 excise 12 internal eliminated sequences that disrupt exons. TPB1 and TPB6 recognize the TIR like sequences. TPB2 and likely a Piggymac in paramecium seem to recognize RNA intermediates. In the case of TPB2, the TIR direct interaction with the TIR has been lost and the domesticated transposase performs its excision via scnRNA-directed heterochromatin involved specialization; thus, TPB2 to recognize heterochromatin rather than TIRs. Thus Tetrahymena TPB2 seemingly represents another level of transposase domestication. The situation is less clear in the case of Piggymac where the domesticated transposase seems to recognize both IES ends and the intermediate RNA. It is likely that in the case of TPB2 and Piggymac, the transposition mechanism was gradually grafted into a heterochromatin formation pathway. In all these cases, the internal eliminated sequence form a circle that prevents reintegration in the genome as this is also the case for jawed vertebrates with the chromosomal excision product resulting from the RAG activity.

The phylogenetic analysis (not shown) indicated that possibly a PiggyBac transposon entered the Oligohymenophorea (that include the paramecium and



**Fig. 14.7** Brief overview of biological and genomic features of the ciliate system. **a** Main steps of the sexual cycle in ciliate species belonging to the *Paramecium* genus. At the entry of the cycle in vegetative cells, the macronucleus starts disorganizing while the 2 micronuclei (<math>2n</math>) trigger meiosis in vegetative cells (1) resulting in eight gametic nuclei (<math>n</math>). Seven or six of these gametic micronuclei, respectively, degenerate after the acquisition of one gametic micronuclei or the exchange of one of them with another *Paramecium* cell. The fusion of both remaining gametic micronuclei (3) leads to a new zygotic nucleus (<math>2n</math>) that thereafter differentiate in new macronuclei and microneclei (4), before karyonidal division and differentiation of a new vegetative cell. **b** Organization of genome sequences in the micronucleus and the macronucleus. Here, the micronucleus genome is reduced to one chromosome containing the genes (blue arrows) that are for some of them split by insertion element sequence (IES; orange rectangle). DNA transposons (orange double arrows) can also be interspersed in chromosomes. During macronucleus differentiation, IES are precisely excised what restores split gene ORFs (the faint blue line means that the 2 ORF are fused in one ORF), transposons are eliminated what fragments chromosomes before the remaining chromosomal fragments are amplified. This figure was constructed from derived information and graphic elements in Nowacki et al. (2011)

tetrahymena phyla) ancestor and became later domesticated. The ancestor Piggymac/TPB2/TPB1/TPB6 domesticated transposon evolved in the two lineages: tetrahymena and paramecium Piggymac and TPB2/TPB1/TPB6. In the case of tetrahymena, a duplication gave rise to TPB2 in one part and TPB1/TPB6 ancestor in the other. TPB2 lost the possibility to recognize its TIR but gained the possibility to interact with the scan RNA. TPB1/TPB6 represent maybe the ancestral state and TPB2 could represent another state in the domestication process.

**Kat 1 and MAT alpha3** (Barsoum et al. 2010; Rajaei et al. 2014)

As we said above, some yeasts have the ability to change their mating type. In *Saccharomyces cerevisiae*, switching is initiated by homothallic switching (HO) endonuclease. In the related yeast *Kluyveromyces lactis* HO has been replaced by two domesticated DDE transposons: MATalpha 3 involved in the switch d *MAT* $\alpha$  to *MATa*. *Kat1* involved in the switch de *MATa* to *MAT* $\alpha$ .

*MAT* $\alpha$ 3 which was domesticated from transposable elements belonging to the Mutator Like Element Family MULEs. Regulated excision of this element results in a double stranded DNA break, (DSB) that stimulates recombination from the genome and initiates mating-type switching from *MAT* $\alpha$  to *MATa*. It has to be noted that in the case of this domesticated transposon the TIR-like sequences have been replaced by other sequences which are different on the left and right sides. The left side is a low complexity sequence with a long stretch of T and A. The right side contains a conserved motif that is conserved in sibling species-species of the *Kluyveromyces* genus. The other domesticated transposon named *Kat1* evolved from hAT (*hobo/Activator/Tam3*) transposases and is involved in the switching from *MATa* to *MAT* $\alpha$ . *Kat1* cleaves the *MATa* locus at two different positions, resulting in DSBs that stimulate recombination. *Kat1* recognizes a TIR like sequence. In both cases, the intervening DNA is joined into a circle.

In conclusion, DDE transposons evolved as site-specific recombination activating genes many times during evolution and therefore this is a case of isoconvergent evolution of site-specific recombination activating genes.

#### ***14.4.2 Simple Evolutionary Shift Can Explain Convergent Evolution***

Isoconvergent evolution, the independent evolution of similar features from the same ancestral state in different evolutionary lineages (Pontarotti and Hue 2016), could be explained in part by natural selection where the new feature gave an advantage to the individual. Isoconvergent evolution could be due to the limited number of evolutionary pathways resulting from developmental and functional constraints on the evolutionary process (Losos 2011)—Functional constraints imposing a finite number of accurate adaptations, a finite number of mechanisms can be used to answer functional problems. Finally, isoconvergent evolution can be explained in part by the

ease of transition from ancestral state to a derived state. This last aspect is not really discussed in the literature.

Losos (2011) pointed out that the wings powering flight in vertebrates have been built in different ways in birds, pterosaurs and bats. In all these cases, the wings represent modified forelimbs. The combination of wings and forelimbs, in theory, would be not very useful in real life. He concluded that this was due to lack of constraints; however, we underline here that besides the constraints, it was easier to modify forelimbs to get wings than to start from nothing. The same reasoning could be applied for the DDE transposon co-option as sequence-specific recombination activating systems and the biochemical shift from a transposase to a sequence-specific recombination activating endonuclease is an easy evolutionary step.

### ***14.4.3 Transposases Form the Largest Family in the Diverse Genomes of Life***

Transposase-encoding genes are greatly over-represented in sequenced genomes and metagenomes relative to other coding sequences (Aziz et al. 2010). Some of the transposase coding genes could correspond to domesticated transposons, active transposons and maybe to fossilized ones. From this and the paragraph developed above, we hypothesize that many DDE transposons have been recruited as recombination activating site-specific endonuclease systems. Many domesticated candidates have already been described in the literature, some of them lost the endonuclease domain but conserved the DNA binding domain and they could be involved in novel chromatin-modifying complexes (Feschotte 2008) while others are involved in centromere binding, chromosome segregation, meiotic recombination (Sinzelle et al. 2009). Some articles also described large-scale systematic analysis to search for domesticated transposons including DDE transposons (Hoen and Bureau 2015; Bouallègue et al. 2017). In the human genome 26 putative DDE transposases have been described (Arnaoty et al. 2012), among these putative domesticated transposases as far as we know other than the RAGs, only the function of one of them, PBGD5 was really investigated, and its nuclease activity has been shown, however we do not know whether it recognizes specific sequences and if it is really involved in recombination (Henssen et al. 2015, 2016, 2017). Another candidate has been tested for its transposase activity (Majumdar et al. 2013) where the authors show that THAP9 gene encodes an active DDE DNA transposase.

We need to test further the hypothesis that many transposons have been recruited as recombination activating site-specific endonuclease. If our hypothesis is true, many domesticated DDE transposons acting as specific DNA endonucleases should be found throughout the life diversity.

#### ***14.4.4 Search for Domesticated Transposons Involved in Programmed Recombination. Test of the Hypothesis***

This paragraph is a small guideline. The strategy to search for domesticated transposons involved in programmed recombination can be performed by two complementary approaches: Look for the candidate domesticated transposase and look for a somatically rearranged genomic region.

**Look for the Candidate Transposase Co-opted as Site-Specific Endonuclease Involved in Genomic Recombination** Different strategies have been developed in particular by Hoen and Bureau (2015) and Bouallègue et al. (2017). The strategy described here was adapted from these publications. We first have to look for the sequence that codes for complete transposase; these can be done by profile search using an alignment with a known transposase (Eddy 2011).

The ORF found needs to be checked for a bona fide catalytic site and the conservation of the DDE motif. Then in order to be sure that the transposase is likely to be active it is necessary to show that the protein evolved under constraint. This can be done by calculating the ratio of non-synonymous to synonymous substitutions (dN/dS) which is a useful measure of the strength and mode of natural selection acting on protein-coding genes (Jeffares et al. 2015).

The next step is to differentiate between transposase belonging to an active transposon from site-specific recombination activating gene (domesticated transposase). The following criteria must be present in the case of site-specific endonuclease. The domesticated transposase must be in single copy without pseudogene like sequences that could correspond to recent transposition events. The next criteria is based on the fact that the domesticated transposase should be unable to transpose thus the domesticated transposase should remain in the same genomic region in different species—conserved synteny—(Rascol et al. 2009). The higher the number of divergent species display conserved synteny, the higher is the probability that the transposase has been domesticated. For example, PGBD5 is present in all chordate genomes and belongs to a conserved synteny (Pavelitz et al. 2013) indicating that it is likely to have been domesticated.

The following step is to search the domesticated sequence recognition signal: find RSS/TIR like sequence. One way to search for the sequence recognition signal is to perform Chip-Seq experiment (Park 2009), and the other, that could be complementary to the Chip-Seq experiment is to be guided by sequence data. This could be done by looking for the active transposon which is the most similar to the domesticated transposon and to determine the TIR sequence of the transposon. The TIR sequence should be similar to the site-specific DNA sequence recognized by the domesticated transposase, as this is the case for the RAG transposon and the domesticated RAG (see Kapitonov and Jurka 2005; Huang et al. 2016; Morales Poole et al. 2017). The TIR like sequences can then be searched in the genome(s) coming from the cells where the “domesticated” transposase is transcribed. It should be noted that the endonuclease could recognize RNA intermediates as this is the case for TPB2 in-



ciliates where the TIR recognition has been lost and instead a chromatin structure is recognized.

**Look for Somatic Rearranged Genomic Regions** The next step is to look for rearranged genomic regions, and this can be done in several ways. The first way would be to sequence the genome from the cell or the tissue where the domesticated transposase is expressed (sequence 1) compare this genome with a reference genome (sequence 2) from the same species (same individual) in order to look for the rearranged sequence in sequence 1. Another way is to use high throughput genome translocation sequencing methodology based on the ability of a double-strand break to translocate to a fixed ‘bait’ double-strand break generated by a nuclease (Hu et al. 2016).

In both cases, the genomic region, where the rearrangement occurred, should be flanked by terminal repeats; these terminal repeats corresponding to the RSS should be compared to the TIR of the active transposon (the most similar one).

Candidate regions using the second methodology have been described by Wei et al. (2016). They described 27 recurrent double-strand breaks that occurred during neural stem/progenitor cell differentiation and they named it as recurrent double-strand break cluster. The authors hypothesized that the break region will be joined to the distal break region thus potentially leading to new genetic information via for example a novel exon combination that can be generated by recombination between intronic regions at the DNA level via “exon shuffling.” (Alt and Schwer 2018) In the case where the double-strand break followed by a recombination occurs via the action of domesticated DDE transposon a possible candidate could be PGBD5. Indeed PGBD5 seems to be expressed in the brain and fetal brain and therefore possibly in the neural stem/progenitor cells (Pavelitz et al. 2013) and PGBD5 is able to induce double double-strand breaks in non-physiologic condition (Henssen et al. 2015, 2017).

**Are There Other Domesticated Transposases that Are Able to Generate Diversity?** One important question concerns the possibility that some domesticated transposase and their TIRs could create diversity through the rearrangement of distinct tandem repeat paralogous gene segments (such as in V(D)J recombination). These processes correspond to an excision between at least 2 sets of paralogous fragments with a combinatorial joining. In these cases, the paralogous fragments should be flanked by the TIR like sequences located 3’ from each of the first paralogous fragments family and 5’ to each of the second fragment paralogous family. Such genomic rearrangement can be searched. The corresponding transposon can be searched as described above.

## 14.5 Conclusion

We discussed in this chapter the possibility that DDE transposons have been recruited several times as systems involved in programmed DNA elimination and perhaps in the generation of receptor diversity. This is due to the fact that the functional shift is an

easy one and that the huge number of transposase domains are present and across the tree of life. Therefore, DDE transposons along with other so-called selfish element encoding nucleases (homing nuclease and prophage) could have been major players in the evolution of biodiversity. The co-option of DDE transposons as the regulatory element is also very important. This has been largely discussed, Justin Goodrich this Issue) (Sinzelle et al. 2009; Jangam et al. 2017). Another important role of DDE transposons is the one in the horizontal gene transfer (HGT). Horizontal transfer of DDE transposons (HTT) has been widely reported in eukaryotes (El Baidouri and Panaud 2015). However, in most cases, the literature discusses HGT and HTT in a non-integrated manner. However, the role of DDE transposons as the carrier of additional antibiotic resistance genes is well known in bacteria (Babakhani and Oloomi 2018), and at least the role of DDE transposon as carrier one case in yeast is also known. (McDonald et al. 2019). It is likely that most of the HGT are driven by transposons, but as most of the studied cases HGT corresponds to ancient events the transposase and the corresponding TIR has been lost so an effort in the analysis of recent HGTs should be done. Therefore, DDE transposons could have been major players in organismal evolution because it helps in genetic exchange between species.

The DDE transposon evolutionary trajectory should be included in the public goods hypothesis for the evolution of life on Earth (McInerney et al. 2011). According to this hypothesis, nucleotide sequences are simply seen as goods, passed from one organism to another through both vertical and horizontal transfer. The interesting things about DDE transposons are that they evolved in vertical and horizontal manner and they should be seen as goods since they increase the possibility of transfer and can be co-opted for example as recombination activating site-specific endonucleases involved in programmed DNA elimination.

**Acknowledgements** This work was supported by the French Government under the «Investissements d'avenir» (Investments for the Future) program managed by the Agence Nationale de la Recherche (ANR, fr: National Agency for Research), (reference: Méditerranée Infection 10-IAHU-03).

## References

- Abe K, Shimizu SY, Tsuda S, Sato T (2017a) A novel non prophage(-like) gene-intervening element within *gerE* that is reconstituted during sporulation in *Bacillus cereus* ATCC10987. *Sci Rep* 7:11426. <https://doi.org/10.1038/s41598-017-11796-8>
- Abe K, Takamatsu T, Sato T (2017b) Mechanism of bacterial gene rearrangement: SprA-catalyzed precise DNA recombination and its directionality control by SprB ensure the gene rearrangement and stable expression of *spsM* during sporulation in *Bacillus subtilis*. *Nucleic Acids Res* 45:6669–6683. <https://doi.org/10.1093/nar/gkx466>
- Alt FW, Schwer B (2018) DNA double-strand breaks as drivers of neural genomic change, function, and disease. *DNA Repair (Amst)* 71:158–163
- Arakawa H, HauschiLd J, Buerstedde JM (2002) Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science* (80-) 295:1301–1306. <https://doi.org/10.1126/science.1067308>

- Arnaoty A, Pitard B, Bateau B, Bigot Y, Lecomte T (2012) Novel approach for the development of new antibodies directed against transposase-derived proteins encoded by human neogenes. *Methods Mol Biol* 859:293–305. [https://doi.org/10.1007/978-1-61779-603-6\\_17](https://doi.org/10.1007/978-1-61779-603-6_17)
- Aziz RK, Breitbart M, Edwards RA (2010) Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* 38:4207–4217. <https://doi.org/10.1093/nar/gkq140>
- Babakhani S, Oloomi M (2018) Transposons: the agents of antibiotic resistance in bacteria. *J Basic Microbiol* 58:905–917
- Barsoum E, Martinez P, Åström SU (2010)  $\alpha 3$ , a transposable element that promotes host sexual reproduction. *Genes Dev* 24:33–44. <https://doi.org/10.1101/gad.557310>
- Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Bétermier M (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev* 23:2478–2483. <https://doi.org/10.1101/gad.547309>
- Boehm T, Hirano M, Holland SJ, Das S, Schorpp M, Cooper MD (2018) Evolution of alternative adaptive immune systems in vertebrates. *Annu Rev Immunol* 36:19–42. <https://doi.org/10.1146/annurev-immunol-042617-053028>
- Bouallègue M, Rouault JD, Hua-Van A, Makni M, Capy P (2017) Molecular evolution of piggyBac superfamily: from selfishness to domestication. *Genome Biol Evol* 9:323–339. <https://doi.org/10.1093/gbe/evw292>
- Cheng CY, Young JM, Lin CYG, Chao JL, Malik HS, Yao MC (2016) The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in *Tetrahymena thermophila*. *Genes Dev* 30:2724–2736. <https://doi.org/10.1101/gad.290460.116>
- Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comput Biol*. <https://doi.org/10.1371/journal.pcbi.1002195>
- El Baidouri M, Panaud O (2015) Horizontal transfers and the new model of TE-driven genome evolution in Eukaryotes. In: *Evolutionary biology: biodiversification from genotype to phenotype*. Springer International Publishing, pp 77–92
- Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA (2015) A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Microbiol* 13:641–650
- Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9:397–405
- Flajnik MF (2016) Evidence of G.O.D.’s miracle: unearthing a RAG transposon. *Cell* 166:11–12
- Henssen AG, Henaff E, Jiang E, Eisenberg AR, Carson JR, Villasante CM, Ray M, Still E, Burns M, Gandara J, Feschotte C, Mason CE, Kentsis A (2015) Genomic DNA transposition induced by human PGBD5. *Elife*. <https://doi.org/10.7554/elife.10565>
- Henssen AG, Jiang E, Zhuang J, Pinello L, Succi ND, Koche R, Gonen M, Villasante CM, Armstrong SA, Bauer DE, Weng Z, Kentsis A (2016) Forward genetic screen of human transposase genomic rearrangements. *BMC Genom* 17:548. <https://doi.org/10.1186/s12864-016-2877-x>
- Henssen AG, Koche R, Zhuang J, Jiang E, Reed C, Eisenberg A, Still E, Macarthur IC, Rodríguez-Fos E, Gonzalez S, Puiggròs M, Blackford AN, Mason CE, De Stanchina E, Gönen M, Emde AK, Shah M, Arora K, Reeves C, Succi ND, Perlman E, Antonescu CR, Roberts CWM, Steen H, Mullen E, Jackson SP, Torrents D, Weng Z, Armstrong SA, Kentsis A (2017) PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet* 49:1005–1014. <https://doi.org/10.1038/ng.3866>
- Hilton JA, Meeks JC, Zehr JP (2016) Surveying DNA elements within functional genes of heterocyst-forming cyanobacteria. *PLoS One*. <https://doi.org/10.1371/journal.pone.0156034>
- Hoen DR, Bureau TE (2015) Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Mol Biol Evol* 32:1487–1506. <https://doi.org/10.1093/molbev/msv042>
- Hu J, Meyers RM, Dong J, Panchakshari RA, Alt FW, Frock RL (2016) Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat Protoc* 11:853–871. <https://doi.org/10.1038/nprot.2016.043>

- Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, Le Petillon Y, Liu X, Chen S, Schatz DG, Xu A (2016) Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* 166:102–114. <https://doi.org/10.1016/j.cell.2016.05.032>
- Jangam D, Feschotte C, Betrán E (2017) Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet* 33:817–831
- Janeway CA, Travers P, Walport M, Shlomchik MJ (2001) Immunobiology: the immune system in health and disease. 5th edn. Garland Science
- Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol Biol* 1201:65–90. [https://doi.org/10.1007/978-1-4939-1438-8\\_4](https://doi.org/10.1007/978-1-4939-1438-8_4)
- Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol* 3:0998–1011. <https://doi.org/10.1371/journal.pbio.0030181>
- Keeling PJ, Roger AJ (1995) The selfish pursuit of sex. *Nature* 375:283
- Klar AJS, Ishikawa K, Moore S (2014) A unique DNA recombination mechanism of the mating/cell-type switching of fission yeasts: a review. *Microbiol Spectr*. <https://doi.org/10.1128/microbiol.spec.mdna3-0003-2014>
- Koufopanou V, Burt A (2005) Degeneration and domestication of a selfish gene in yeast: molecular evolution versus site-directed mutagenesis. *Mol Biol Evol* 22:1535–1538. <https://doi.org/10.1093/molbev/msi149>
- Losos JB (2011) Convergence, adaptation, and constraint. *Evolution (N Y)* 65:1827–1840. <https://doi.org/10.1111/j.1558-5646.2011.01289.x>
- Majumdar S, Singh A, Rio DC (2013) The human THAP9 gene encodes an active P-element DNA transposase. *Science (80-)* 339:446–448. <https://doi.org/10.1126/science.1231789>
- McDonald MC, Taranto AP, Hill E, Schwessinger B, Liu Z, Simpfendorfer S, Milgate A, Solomon PS (2019) Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. *MBio*. <https://doi.org/10.1128/mbio.01515-19>
- McInerney JO, Pisani D, Baptiste E, O'Connell MJ (2011) The public goods hypothesis for the evolution of life on Earth. *Biol Direct* 6:41. <https://doi.org/10.1186/1745-6150-6-41>
- Morales Poole JR, Huang SF, Xu A, Bayet J, Pontarotti P (2017) The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates. *Immunogenetics* 69:391–400. <https://doi.org/10.1007/s00251-017-0979-5>
- Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, Landweber LF (2009) A functional role for transposases in a large eukaryotic genome. *Science* 324:935–938
- Nowacki M, Shetty K, Landweber LF (2011) RNA-mediated epigenetic programming of genome rearrangements. *Annu Rev Genomics Hum Genet* 12:367–389
- Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10:669–680
- Pavelitz T, Gray LT, Padilla SL, Bailey AD, Weiner AM (2013) PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mob DNA* 4:23. <https://doi.org/10.1186/1759-8753-4-23>
- Piégu B, Bire S, Arensburger P, Bigot Y (2015) A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol Phylogenet Evol* 86:90–109
- Pontarotti P, Hue I (2016) Road map to study convergent evolution: A proposition for evolutionary systems biology approaches. *Evolutionary biology: convergent evolution, evolution of complex traits, concepts and methods*. Springer International Publishing, Heidelberg, pp 3–21
- Rajaei N, Chiruvella KK, Lin F, Åström SU (2014) Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc Natl Acad Sci U S A* 111:15491–15496. <https://doi.org/10.1073/pnas.1406027111>

- Rascol VL, Levasseur A, Chabrol O, Grusea S, Gouret P, Danchin EGJ, Pontarotti P (2009) CASSIOPE: an expert system for conserved regions searches. *BMC Bioinf* 10:284. <https://doi.org/10.1186/1471-2105-10-284>
- Sinzelle L, Izsvák Z, Ivics Z (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* 66:1073–1093
- Teng G, Schatz DG (2015) Regulation and evolution of the RAG recombinase. In: *Advances in immunology*. Academic Press Inc, pp 1–39
- Wang J, Davis RE (2014) Programmed DNA elimination in multicellular organisms. *Curr Opin Genet Dev* 27:26–34
- Wei PC, Chang AN, Kao J, Du Z, Meyers RM, Alt FW, Schwer B (2016) Long neural genes harbor recurrent DNA break clusters in neural stem/progenitor cells. *Cell* 164:644–655. <https://doi.org/10.1016/j.cell.2015.12.039>
- Yu K, Lieber MR (2019) Current insights into the mechanism of mammalian immunoglobulin class switch recombination. *Crit Rev Biochem Mol Biol* 54:333–351
- Zhang Y, Cheng TC, Huang G, Lu Q, Surleac MD, Mandell JD, Pontarotti P, Petrescu AJ, Xu A, Xiong Y, Schatz DG (2019) Transposon molecular domestication and the evolution of the RAG recombinase. *Nature* 569:79–84. <https://doi.org/10.1038/s41586-019-1093-7>
- Zufall RA, Robinson T, Katz LA (2005) Evolution of developmentally regulated genome rearrangements in eukaryotes. *J Exp Zool Part B Mol Dev Evol* 304:448–455

# Chapter 15

## Evolution of Milk Oligosaccharides of Carnivora and Artiodactyla: Significance of the Ratio of Oligosaccharides to Lactose in Milk



Tadasu Urashima, Yuri Mineguchi, Kenji Fukuda, Katherine Whitehouse-Tedd, and Olav T. Oftedal

**Abstract** Mammalian milk and colostrum usually contain lactose as a predominant saccharide as well as lower concentrations of a variety of milk oligosaccharides. However, in the milk of monotremes and marsupials, oligosaccharides predominate over lactose. Among eutherians, many species of the order Carnivora are also exceptional in that they contain substantial amounts of oligosaccharides in addition to lactose. With the exception of the domestic dog, milk oligosaccharides predominate over lactose in the milk of Caniformia, including mink, striped skunk, raccoon, many bears and seals, whereas lactose is the dominant saccharide in the milk of some species of Feliformia, such as spotted hyena, African lion, clouded leopard and cheetah. A significant feature of the milk oligosaccharides of Carnivora is the presence of A (GalNAc $\alpha$ 1-3(Fuc( $\alpha$ 1-2)Gal), B (Gal $\alpha$ 1-3(Fuc $\alpha$ 1-2)Gal) or H (Fuc $\alpha$ 1-2Gal) units as well as  $\alpha$ -Gal (Gal $\alpha$ 1-3Gal) unit, attached to the core structures of lactose, lacto-*N*-neotetraose (Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc), lacto-*N*-neohexaose (Gal $\beta$ 1-4GlcNAc $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc) or para lacto-*N*-neohexaose (Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc). The presence of A, B, H antigens, and  $\alpha$ -Gal varies depending on each species of Carnivora. In contrast to the milk of Carnivora, that of Artiodactyla including such diverse species as giraffe, sitatunga, deer and water buffalo, contains lactose as the predominant saccharide, although small amounts of oligosaccharides are present as well. In most of these oligosaccharides, the core structure is lactose and they all contain isoglobotriose (Gal $\alpha$ 1-3Gal $\beta$ 1-4Glc), but are heterogeneous with respect to the presence of a few saccharides such as GM2 tetrasaccharide (Neu5Ac $\alpha$ 2-3(GalNAc $\beta$ 1-4)Gal $\beta$ 1-4Glc) and globotriose (Gal $\alpha$ 1-4Gal $\beta$ 1-4Glc). None of these milks include

---

T. Urashima (✉) · Y. Mineguchi · K. Fukuda  
Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Japan  
e-mail: [urashima@obihiro.ac.jp](mailto:urashima@obihiro.ac.jp)

K. Whitehouse-Tedd  
School of Animal, Rural and Environmental Sciences, Nottingham Trent University,  
Southwell, Nottinghamshire, UK

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

oligosaccharides containing A, B or H antigens. In addition, milk or colostrum of other Artiodactyla such as cow, sheep, goat, camel and pig contain very low concentrations of oligosaccharides whose core structures are lacto-*N*-neotetraose, lacto-*N*-neohexaose, lacto-*N*-novopentaose I (Gal $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc( $\beta$ 1-6)Gal $\beta$ 1-4Glc) or GlcNAc $\beta$ 1-3 (Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc. In this chapter, we hypothesize on the evolution of milk oligosaccharides in Carnivora and Artiodactyla, as well as on the potential significance of the ratio of oligosaccharides to lactose in these milks.

## Abbreviations

Glc	Glucose
Gal	Galactose
GlcNAc	<i>N</i> -acetylglucosamine
GalNAc	<i>N</i> -acetylgalactosamine
Fuc	Fucose
Neu5Ac	<i>N</i> -acetylneuraminic acid
Neu5Gc	<i>N</i> -glycolylneuraminic acid
OS	Oligosaccharide

## 15.1 Introduction

Although mammalian milk or colostrum usually contain lactose as a predominant saccharide as well as lesser concentrations of many varieties of milk oligosaccharides (Jenness et al. 1964; Messer and Urashima 2002; Urashima et al. 2014), oligosaccharides predominate over lactose in the milks of monotremes and marsupials (Messer and Urashima 2002; Urashima et al. 2014; Urashima and Messer 2017). It has been hypothesized that in suckling monotremes and marsupials, the milk oligosaccharides are absorbed in the small intestine by pinocytosis or endocytosis and hydrolyzed by lysosomal enzymes, the resulting monosaccharides being transferred into circulation and then utilized as energy sources (Messer and Urashima 2002; Urashima et al. 2014; Urashima and Messer 2017). Among most eutherian neonates, lactose is hydrolyzed to glucose and galactose by small intestinal lactase and these monosaccharides are then absorbed and enter circulation. Glucose is directly utilized as an energy source, while most of the galactose is converted to glucose in the liver in order to be utilized as an energy source. Thus, lactose is thought to be a significant energy source for most eutherian neonates (Messer and Urashima 2002; Urashima et al. 2014).

It is well recognized, however, that in human infants most of the milk oligosaccharides are hydrolyzed partially if at all in the small intestine and thus intact oligosaccharides reach the colon. Recent evidence suggests that some human milk oligosaccharides (HMOs) act as prebiotics that stimulate the growth of beneficial colonic

bacteria, and some may act as anti-infection agents against pathogenic bacteria or viruses, or as immune-modulation and anti-inflammation agents, and some provide prevention against enterocolitis and assist recovery of the colonic barrier function, and stimulate brain activity (Bode 2012). The anti-microbial action of HMOs against pathogens has mostly been studied in terms of inhibition of microorganism adhesion to the host in in vitro investigations using epithelial cells, but a few studies for anti-infection, including against *Campylobacter jejuni* and enteropathogenic *Escherichia coli*, have been done by in vivo experiment with a mice model (Yu et al. 2016; He et al. 2016). It was recently shown that the growth of group B Streptococcus (GBS), which causes meningitis to the host, was inhibited by the addition of an HMOs mixture during in vitro incubation (Lin et al. 2017; Craft et al. 2018).

Human milk contains 12–13 g/L of milk oligosaccharides, which constitute the third-largest solid component after lactose and lipid in the milk. To date, more than 240 varieties of HMOs have been separated from human milk, of which 169 have been characterized (Urashima et al. 2018a). These are classified into 20 series based on their core structures. In the 169 HMOs, fucose (Fuc) or N-acetylneuraminic acid (Neu5Ac) residues are linked to diverse positions in the structures of galactose (Gal), N-acetylglucosamine (GlcNAc) or glucose (Glc).

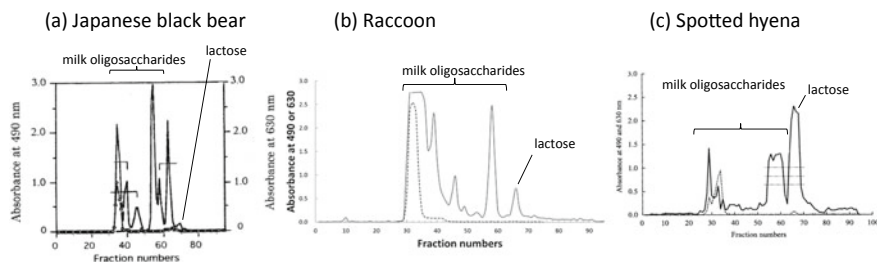
Among eutherian species, the Carnivora, especially Caniformia species are exceptions in that oligosaccharides predominate over lactose in their milk, whereas the milk or colostrum of the Artiodactyla species contains lactose as the dominant saccharides in addition to lesser concentrations of oligosaccharides, similar to most other eutherian species. In this chapter, we compare the milk oligosaccharide structures between the Carnivora and the Artiodactyla, discuss their evolution in both orders and hypothesize on the factors that determine the ratio of milk oligosaccharides to lactose.

## 15.2 Biochemical Properties and Characterization of Milk Oligosaccharides in Carnivora Species

The order Carnivora consists of two suborders, Caniformia including Canidae (e.g., dogs, foxes, wolves), Ursidae (bears), Phocidae (true seals), Otariidae (fur seals), Odobenidae (walrus), Ailuridae (red panda), Mephitidae (striped skunk), Procyonidae (raccoons, coatis), Mustelidae (mink, weasels) and Feliformia including Felidae (cats, lion, clouded leopard, cheetah), Viverridae (civet), Hyaenidae (hyena), Herpestidae (mongoose), etc.

Figure 15.1 shows the gel filtration profiles of carbohydrate fractions separated from milk of Japanese black bear (Ursidae) (Urashima et al. 1999a), raccoon (Procyonidae) (Urashima et al. 2018b) and spotted hyena (Hyaenidae) (Uemura et al. 2009), indicating the peaks of lactose and milk oligosaccharides as shown in Figs. 15.1a–c. The peak of lactose is due to a single component, while the milk oligosaccharides,

















**Fig. 15.1** Profiles of milk carbohydrate of **a** Japanese black bear, **b** raccoon and **c** spotted hyena by gel filtration on BioGel P-2 column (2.5 × 100 cm). Solid line: detection at 490 nm by phenol—H<sub>2</sub>SO<sub>4</sub> method. Dotted line: detection at 630 nm by periodate—resorcinol method. The chromatograms were from **a** Urashima et al. (1999a), **b** Urashima et al. (2018b) and **c** Uemura et al. (2009)

which elute earlier than lactose, correspond to multiple peaks. Milk oligosaccharides predominate over lactose in milks of Japanese black bear and raccoon, whereas lactose is the predominant saccharide in milk of spotted hyena (Fig. 15.1). The ratio of milk oligosaccharides (milk OS) to lactose was estimated from the peak area with the solid line for hexose, the absorbance of which was detected by the phenol—H<sub>2</sub>SO<sub>4</sub> method. The neutral and acidic oligosaccharides were separated by anion exchange chromatography and finally purified by high performance liquid chromatography (HPLC) with a graphite carbon column and a reverse phase system, respectively. Each purified oligosaccharide was characterized by proton nuclear magnetic resonance spectroscopy (<sup>1</sup>H-NMR) and matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS).

The neutral and acidic milk oligosaccharides of Carnivora species were characterized for dog (Bubb et al. 1999; Rostami et al. 2014), mink (Urashima et al. 2005), striped skunk (Taufik et al. 2013), raccoon (Urashima et al. 2018b), white-nosed coati (Urashima et al. 1999b), Japanese black bear (Urashima et al. 1999a, 2004), American black bear (Urashima et al. 2019), polar bear (Urashima et al. 2000), brown bear (Urashima et al. 1997), giant panda (Nakamura et al. 2003a), hooded seal (Urashima et al. 2001), harbor seal (Urashima et al. 2003), spotted hyena (Uemura et al. 2009), African lion (Senda et al. 2010), clouded leopard (Senda et al. 2010) and cheetah (Urashima et al. 2019). The main neutral milk oligosaccharides in these species are shown in Table 15.1. The milk oligosaccharides are compared among these species according to their characterized oligosaccharide structures. 2'-Fucosyllactose (Fucα1-2Galβ1-4Glc) was found in the milks of many species other than giant panda, clouded leopard and cheetah, but its concentration was low in species whose milk contained A or B tetrasaccharides. Isoglobotriose (Galα1-3Galβ1-4Glc) was found in carnivore milks other than dog, raccoon, hooded and harbor seals, African lion and clouded leopard. A-tetrasaccharide (GalNAcα1-3(Fucα1-2)Galβ1-4Glc) was identified in the milks of dog, striped skunk, polar bear, African lion and clouded leopard, while B-tetrasaccharide (Galα1-3(Fucα1-2)Galβ1-4Glc) was found in the milks of Japanese black bear, American black bear, polar

**Table 15.1** Comparison of milk oligosaccharides among Carnivora species

	Dog	Mink	Striped skunk	Raccoon	White-nosed coati	Japanese black bear	Polar bear	Brown bear	Giant panda	Hooded seal	Spotted hyena	African lion	Clouded leopard	Cheetah
a 	+	+	+	+	+	+	+	+	-	+	+	+	-	-
b 	-	-	-	+	+	-	-	-	-	+	-	-	-	-
c 	-	+	+	-	+	+	+	+	+	-	+	-	-	+
d 	-	-	+	-	+	-	+	-	-	-	-	-	-	-
e 	+	-	+	-	-	-	+	-	-	-	-	+	+	-
f 	-	-	-	-	-	+	+	-	-	-	+	-	-	-
g 	-	-	-	-	-	-	-	+	-	-	-	-	-	-
h 	-	-	-	-	-	+	+	-	+	-	-	-	-	-
i 	-	-	-	-	-	-	+	+	-	-	-	-	-	-
j 	-	-	-	-	-	-	+	-	-	-	-	-	-	-
k 	-	-	-	-	-	+	-	-	-	-	-	-	-	-
l 	-	-	-	-	-	+	-	-	-	-	-	-	-	-

a: Fuca1-2Galβ1-4Glc  
 b: Fuca1-2Galβ1-4GlcNAcβ1-3Galβ1-4Glc  
 c: Galα1-3Galβ1-4Glc  
 d: Galα1-3Galβ1-4GlcNAcβ1-3Galβ1-4Glc  
 e: GalNAca1-3(Fuca1-2)Galβ1-4Glc  
 f: Galα1-3(Fuca1-2)Galβ1-4Glc  
 g: Fuca1-2Galβ1-4(Fuca1-3)GlcNAcβ1-3Galβ1-4Glc  
 h: Galα1-3Galβ1-4(Fuca1-3)glc  
 i: Galα1-3Galβ1-4(Fuca1-3)GlcNAcβ1-3Galβ1-4Glc  
 j: GalNAca1-3(Fuca1-2)Galβ1-4(Fuca1-3)Glc  
 k: Galα1-3(Fuca1-2)Galβ1-4(Fuca1-3)Glc  
 l: Galα1-3(Fuca1-2)Galβ1-4(Fuca1-3)GlcNAcβ1-3Galβ1-4Glc

bear and spotted hyena. A-pentasaccharide (GalNAca1-3(Fuca1-2)Galβ1-4(Fuca1-3)Glc) was found in the milk of polar bear, while B-pentasaccharide (Galα1-3(Fuca1-2)Galβ1-4(Fuca1-3)Glc) was identified in the milk of Japanese black bear and American black bear. Among these Carnivora species, only the milks of bears and giant panda contained Lewis x (Galβ1-4(Fuca1-3)Glc(NAc)) containing oligosaccharides. (Oligosaccharides of American black bear and harbor seal are not listed in Table 15.1, due to space limitations.)

Comparison of the core structures of milk oligosaccharides among the Carnivora species shows that all had oligosaccharides containing a lactose core unit. Species other than dog, giant panda, spotted hyena, African lion and clouded leopard also had oligosaccharides containing lacto-*N*-neotetraose (Galβ1-4GlcNAcβ1-3Galβ1-4Glc) or lacto-*N*-neohexaose (Galβ1-4GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc) as core units, while oligosaccharides containing para lacto-*N*-neohexaose (Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4Glc) were found only in the milks of raccoon and hooded seal.

It can be concluded that a characteristic feature of the neutral milk oligosaccharides of Carnivora is the presence of A (GalNAca1-3(Fuca1-2)Gal), B (Galα1-3(Fuca1-2)Gal) or H (Fuca1-2Gal) antigens as well as α-Gal (Galα1-3Gal) attached

to the core structures of lactose, lacto-*N*-neotetraose (Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc), lacto-*N*-neohexaose (Gal $\beta$ 1-4GlcNAc $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc) or para lacto-*N*-neohexaose (Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc). The presence of A, B, H antigens, and  $\alpha$ -Gal vary depending on each species of Carnivora.

The acidic milk oligosaccharides can be compared among dog, mink, striped skunk, raccoon, Japanese black bear, American black bear, giant panda, harbor seal, spotted hyena, African lion, clouded leopard and cheetah. Among these, the milk of species other than African lion and clouded leopard contained the oligosaccharides containing Neu5Ac, while the milks of these two species had Neu5Gc $\alpha$ 2-3Gal $\beta$ 1-4Glc. Neu5Ac $\alpha$ 2-3Gal $\beta$ 1-4Glc (3'-SL) was identified in the milks of dog, striped skunk, Japanese black bear, American black bear, giant panda and spotted hyena but not in those of mink, raccoon, harbor seal, lion, clouded leopard, while the larger oligosaccharides containing lacto-*N*-neohexaose unit and  $\alpha$ 2-3 linked Neu5Ac were identified only in raccoon milk. In cheetah milk, 3'-SL was not identified, but Neu5Ac $\alpha$ 2-8Neu5Ac $\alpha$ 2-3Gal $\beta$ 1-4Glc (DSL) was found. Neu5Ac $\alpha$ 2-6Gal $\beta$ 1-4Glc (6'-SL) was found in the milks only of dog and giant panda, while larger oligosaccharides containing lacto-*N*-neohexaose as well as  $\alpha$ 2-6 linked Neu5Ac were identified in the milks of mink, striped skunk, raccoon, Japanese black bear, American black bear and harbor seal. Among these species, only dog and cheetah milk contained Gal $\beta$ 1-4Glc-3'-O-sulfate (lactose sulfate).

The ratio of oligosaccharides to lactose in the milks of the Carnivora species was estimated from the peak areas in the profiles of the carbohydrate fractions on gel filtration, as shown in Table 15.2. It is clear that milk oligosaccharides predominate over lactose in the milks of the Caniformia species other than dog, especially in those of Japanese black bear, American black bear and raccoon, whereas in the milks of the Feliformia species the concentration of oligosaccharides is fairly similar to that of lactose.

**Table 15.2** Ratio of oligosaccharides to lactose in milks among some Carnivora and Artiodactyla species

	Dog	Mink	Striped skunk	Raccoon	White-nosed coati	Japanese black bear	American black bear	Polar bear	Giant panda	Harbour seal	Spotted hyena	African lion	Clouded leopard	Cheetah
Milk OS : lactose	1:6	5:1	7:1	32:1	2:1	52:1	21:1	13:1	10:1	4:1	1:1	1:2	1:1	1:1

	Dromedary camel mature milk	Bactrian camel colostrum	Bactrian camel mature milk	Giraffe mature milk	Sitatunga colostrum	Addax colostrum	Deer mature milk	Reindeer mature milk	Water buffalo colostrum	Cow colostrum	Cow mature milk	Sheep colostrum
Milk OS : lactose	1:6	1:1.6	1:11	1:12	1:4.2	1:2.8	1:7	1:4.5	1:5	1:3	1:22	1:13

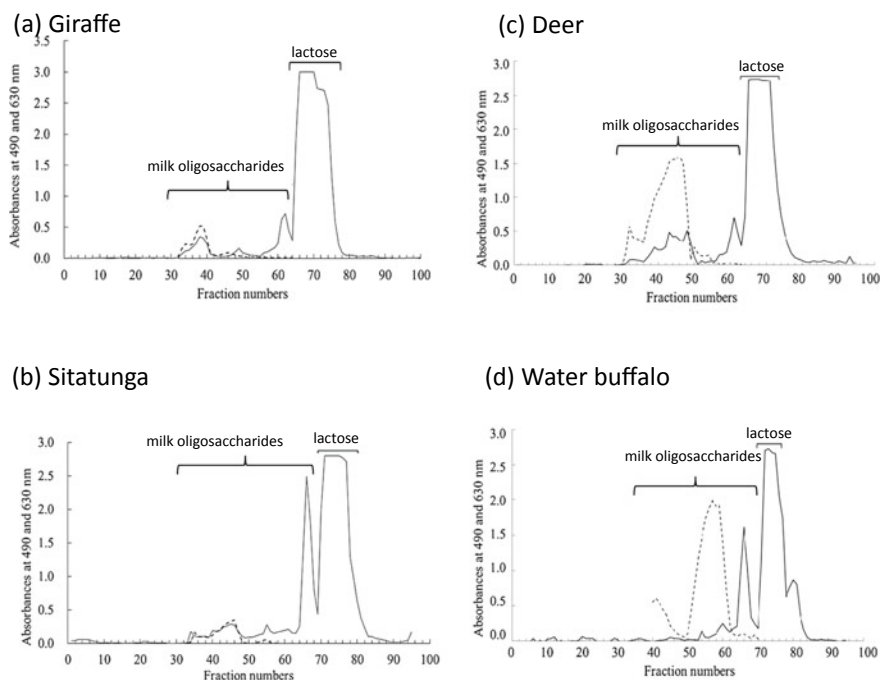
One can hypothesize about why milk oligosaccharides predominate over lactose in milks of both bears and seals. Species in both taxa rely heavily on stored reserves of lipid and protein as sources of substrate for milk production: The mammary gland demand for glucose is minimized by low sugar (lactose and oligosaccharides) in milk. In fact most bears give birth in dens or brush piles, and lactate while hibernating for several months without eating or drinking. In general, lactose is a major osmolyte and is responsible for osmotic movement of water into mammary milk, and thus increased milk water content; however, oligosaccharides have a much lower osmotic effect (per gram). We hypothesize that high oligosaccharides: lactose milks evolved in bears not only as a means of providing diverse saccharides to denned cubs, but also to reduce the water demand of lactation in non-drinking, hibernating mothers. The extent to which water concentration is as important in lactating marine seals is less certain, as is the relationship of oligosaccharides: lactose to the tremendous range of milk fat and dry matter among seals species. Female harbor seals have rich stores of subcutaneous fat and it is thought that their cubs have little need for carbohydrate and depend on milk fat to provide insulating material as adipose tissue (blubber) and as an energy source.

In addition, it has been observed that in milk of the giant panda the concentrations of oligosaccharides change dramatically during the course of lactation. These oligosaccharides are mainly sialyllactose (probably 3'-SL), the concentration of which fell until 20–30 days, and fucosyllactose (either 2'-fucosyllactose or Gal $\beta$ 1-4(Fuc $\alpha$ 1-3)Glc (3-fucosyllactose)), which began to increase at that time (Griffiths et al. 2015). Notably, although the giant panda is related to bears (Ursidae), it does not undergo hibernation.

It seems likely that since the cubs and pups of bears and seals do not depend on milk carbohydrate to supply their energy, the expression level of  $\alpha$ -lactalbumin, which is an essential subunit of lactose synthase of the lactating mammary glands, would be low (Messer and Urashima 2002; Urashima et al. 2012). This could result in slow biosynthesis of lactose but nevertheless be sufficient to permit the biosynthesis of lactose-containing oligosaccharides (Messer and Urashima 2002; Urashima et al. 2012).

### 15.3 Biochemical Properties and Characterization of Milk Oligosaccharides in the Artiodactyla Species











Until recently, the milk oligosaccharides of the Artiodactyla were only characterized for domestic farm animal species including cows, goats, sheep, camels and pigs. However, the oligosaccharides separated from the milks of giraffe, sitatunga, deer and water buffalo have now been characterized (Mineguchi et al. 2018). The profiles of the carbohydrate fractions separated from the milk or colostrum of these species using gel filtration are shown in Fig. 15.2a–d, indicating that lactose is the predominant saccharide along with lower concentrations of milk oligosaccharides as similar



**Fig. 15.2** Profiles of milk carbohydrates of **a** giraffe, **b** sitatunga, **c** deer and **d** water buffalo by gel filtration on BioGel P-2 column (2.5 × 100 cm). Solid line: detection at 490 nm by phenol—H<sub>2</sub>SO<sub>4</sub> method. Dotted line: detection at 630 nm by periodate—resorcinol method

in milks/colostra of the above domestic species. This means that milk/colostrum of not only the domestic Artiodactyla but also the non-domestic species contains lactose as a dominant carbohydrate with lesser concentrations of oligosaccharides. Each oligosaccharide was separated from the peak fractions by reverse phase HPLC and characterized by <sup>1</sup>H-NMR and MALDI-TOF MS. The identified oligosaccharides are shown in Table 15.3. Their core structure was lactose except for lacto-*N*-neotetraose in deer milk. Isogobotriose was found in all these milks/colostra, while Gal $\alpha$ 1-4Gal $\beta$ 1-4Glc (globotriose) and Neu5Ac $\alpha$ 2-3(GalNAc $\beta$ 1-4)Gal $\beta$ 1-4Glc (GM2 tetrasaccharide) were identified only in milk/colostrum of sitatunga and giraffe, respectively. Gal $\beta$ 1-3Gal $\beta$ 1-3Gal $\beta$ 1-4Glc (digalactosyllactose), which was found in water buffalo milk, had up to date, been identified only in marsupial milk (Urashima et al. 2014, 2017). Lactose sulfate was found only in deer milk. 3'-SL, 6'-SL, Gal $\beta$ 1-3Gal $\beta$ 1-4Glc (3'-GL) and Gal $\beta$ 1-6Gal $\beta$ 1-4Glc (6'-GL), which were found in either milk or colostrum, have been identified in the milk or colostrum of many other species, including cows, goats, sheep, camels and pigs (Albrecht et al. 2014). The ratio of milk oligosaccharides to lactose in milk or colostrum was estimated by the profiles of gel filtrations as shown in Table 15.2. Table 15.2 also includes this ratio in milk or colostrum of cow (Fukuda et al. 2010), sheep (Sasaki et al. 2016), reindeer

**Table 15.3** Comparison of milk oligosaccharides among giraffe, sitatunga, deer and water buffalo

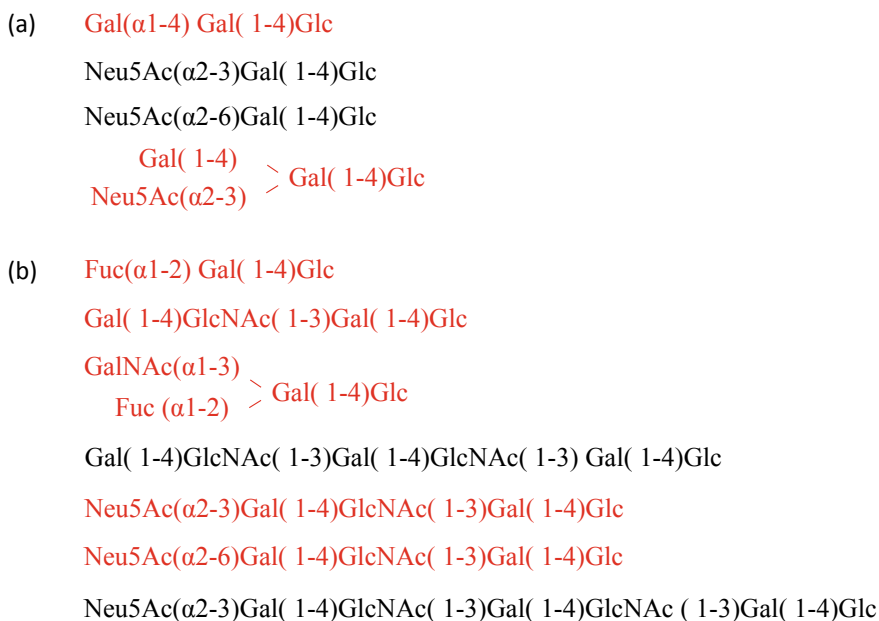
Structure	CFG format	giraffe	sitatunga	deer	water buffalo
Gal $\alpha$ 1-3Gal $\beta$ 1-4Glc		+	+	+	+
Gal $\alpha$ 1-4Gal $\beta$ 1-4Glc			+		
Gal $\beta$ 1-3Gal $\beta$ 1-4Glc				+	+
Gal $\beta$ 1-6Gal $\beta$ 1-4Glc				+	
Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc				+	
Gal $\beta$ 1-3Gal $\beta$ 1-3Gal $\beta$ 1-4Glc					+
Gal $\beta$ 1-4Glc-3'-O-S				+	
Neu5Ac $\alpha$ 2-3Gal $\beta$ 1-4Glc			+		
Neu5Ac $\alpha$ 2-6Gal $\beta$ 1-4Glc			+		
GalNAc $\beta$ 1-4Gal $\beta$ 1-4Glc Neu5Ac $\alpha$ 2-3		+			

(Taufik et al. 2014), addax (Ganzorig et al. 2018), Bactrian camel (Fukuda et al. 2010) and dromedary camel (Alhaj et al. 2013). The predominance of lactose over milk oligosaccharides among the Artiodactyla species differs clearly from that of the Carnivora, especially in Caniformia, but resembles that of most eutherian species.

Even though the core structure of these oligosaccharides was only lactose, except for lacto-*N*-neotetraose in deer milk, it is possible that these milks or colostrum also contained small concentrations of unidentified oligosaccharides with other core structures. A limitation of the use of <sup>1</sup>H-NMR for characterizing oligosaccharides is that it requires rather large sample amounts. Albrecht et al. (2014) identified the milk oligosaccharides of domestic farm artiodactyls, i.e., cows, goats, sheep, dromedary camels and pigs using HPLC with a hydrophilic interaction column, successive exoglycosidase digestions as well as mass spectrometry. Their method required only small amounts of the samples for characterization. It was found that these milks or colostrum contained the oligosaccharides with core units of lacto-*N*-neotetraose, lacto-*N*-neohexaose, lacto-*N*-novopentaose I (Gal $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc) or GlcNAc $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc as well as lactose. If one use this method to characterize the oligosaccharides in milk or colostrum of non-domestic Artiodactyla such as giraffe and sitatunga, it would be possible to find the oligosaccharides containing such core structures. Fukuda et al. (2010) identified Gal $\beta$ 1-3(Neu5Ac $\alpha$ 2-6Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc, Neu5Ac $\alpha$ 2-3Gal $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc and Neu5Ac $\alpha$ 2-6Gal $\beta$ 1-4GlcNAc $\beta$ 1-3(Gal $\beta$ 1-4GlcNAc $\beta$ 1-6)Gal $\beta$ 1-4Glc in Bactrian camel colostrum, while Urashima et al. (1991) found lacto-*N*-novopentaose 1 in bovine colostrum as well. These results show that the Artiodactyla milks/colostrum contain oligosaccharides with the above core units as well as lactose; the lactose core unit is predominant. It appears that among species whose milk or colostrum contains lactose as a predominant saccharide, oligosaccharides that contain a lactose core unit predominate over oligosaccharides that contain

other core units, such as lacto-*N*-neotetraose and lacto-*N*-neohexaose. This is the case for these Artiodactyla milks or colostrum.

It is well recognized that the Artiodactyla is phylogenetically close to the Cetacea and classified within Cetartiodactyla. This suggests that there may be some homologies in milk oligosaccharides between the two orders. Uemura et al. (2005) studied the oligosaccharides in the milk of bottlenose dolphin, a toothed whale, and identified globotriose, GM2 tetrasaccharide, 3'-SL and 6'-SL as shown in Fig. 15.3a. As GM2 tetrasaccharide and globotriose have also been found in the milk/colostrum of giraffe and sitatunga, respectively, this may suggest homology in the milk oligosaccharides between bottlenose dolphin and these two Artiodactyla species. Urashima et al. (2002) identified the oligosaccharides in the milk of the minke whale, one of the baleen whales, as shown in Fig. 15.3b. As Fuc $\alpha$ 1-2Gal $\beta$ 1-4Glc (2'-FL), lacto-*N*-neotetraose, A-tetrasaccharide, Neu5Ac $\alpha$ 2-3Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc and Neu5Ac $\alpha$ 2-6Gal $\beta$ 1-4GlcNAc $\beta$ 1-3Gal $\beta$ 1-4Glc (LST c) have also been found in bovine colostrum, this suggests that there may be homologies in milk oligosaccharides between both Cetartiodactyla species. However, the number of milk oligosaccharides in Cetacean species that have been characterized so far is still too small



**Fig. 15.3** Milk oligosaccharides of **a** bottlenose dolphin and **b** minke whale. **a** The structures written in red have also been identified in milks of sitatunga or giraffe (Mineguchi et al. 2018; Uemura et al. 2005). **b** The structures written in red have also been identified in bovine colostrum (Urashima et al. 2002; Albrecht et al. 2014)

to permit definitive conclusions with respect to the homology or heterogeneity of milk oligosaccharides between Cetacean and Artiodactyla. The milk carbohydrates have been studied only for bottlenose dolphin (Uemura et al. 2005), Minke whale (Urashima et al. 2002), beluga (Urashima et al. 2002), Bryde's whale (Urashima et al. 2007) and Sei whale (Urashima et al. 2007) among the Cetacea.

## 15.4 Physiological Significance of the Ratio of Oligosaccharides to Lactose in Milk

### 15.4.1 *Carnivora*

The Carnivora (and especially Caniformia other than domestic dog) is the only one among many orders of eutherians in which oligosaccharides predominate over lactose in milk or colostrum. The ratio of milk OS to lactose as shown in Table 15.2 illustrates that oligosaccharides predominate over lactose in milks of Caniformia species, but not in milks of Feliformia. However, a complication is that among different mammals the milk OS to lactose may also differ at different lactation stages after birth. In our survey, milk or colostrum samples were collected at the following lactation stages; domestic dog at 13 days postpartum (pp), mink at 15 days pp, striped skunk at 20–48 days pp, white-nosed coati at 17 days pp, Japanese black bear at 37 days pp, American black bear at 0–2 days pp, polar bear at 27 months pp, giant panda at 13 days pp, spotted hyena at 2 days pp, African lion at 127 days pp and cheetah at 0–2 days pp. The change in this ratio during the course of lactation is evident in human milk. The OS concentration decreases over time (i.e., 22–24 g/L in colostrum and 12–13 g/L in mature milk, respectively) (Bode 2012) with an OS:lactose ratio of around 1:2.6 in colostrum and 1:4.6 in mature milk. The similar changes may be anticipated to occur in other species. In case of cows, colostrum contains more than 1 g/L of sialyl oligosaccharides, but this decreases to trace level after 48 h postpartum (Nakamura et al. 2003b). The ratio of milk OS to lactose was 1:3 in colostrum and 1:22 in mature milk (see Table 15.2). In case of Bactrian camels, this ratio was 1:1.6 in colostrum and 1:11 in mature milk (see Table 15.2). This indicates that the ratio may vary during the course of lactation and that inter-specific comparisons should be conducted with milk samples from the same lactation stages.

Although the biological significance of the differing proportions of lactose, acidic and neutral oligosaccharides observed among mammals is not certain, we discussed it in our published paper as described in the following (Urashima et al. 2020). The hypothesis that oligosaccharides serve an anti-bacterial and prebiotic function in neonatal digestive tracts, and thus should be selected for in social species with larger group size and more avenues for social transmission of pathogens (Tao et al. 2011) does not appear to apply to carnivorans. Highly social lions and dogs have the lowest,



not highest, OS:lactose ratios, whereas high OS:lactose ratios are found in predominantly solitary carnivorans, such as raccoons, bears, mink and striped skunk (see above).

An alternative hypothesis is that oligosaccharides, as transporters of specific saccharide constituents, may be important for species with altricial young. Certainly, altricial neonates are characterized by immaturity of physiological and biochemical functions, and when this is combined with rapid growth and the need to rapidly synthesize complex tissues, a situation could arise in which rates of tissue synthesis of a particular constituent are insufficient to meet requirements without dietary (i.e., milk) supply. For example, in altricial rat pups, the rate-limiting enzyme in sialic acid synthesis (UDP-*N*-acetylglucosamine *N*-acetylmannosamine epimerase/kinase) has low liver activity in neonates (Gal et al. 1997). Based on gene expression profiles of various tissues of rat pups, Duncan et al. (2009) concluded that sialic acid was absorbed from milk and compensated for low neonatal rates of sialic acid synthesis. Radiolabelled exogenous sialic acid has also been shown to be deposited in the tissues of growing rat pups, including the brain, which has an especially high sialic acid content (Wang et al. 1998; Sprenger and Duncan 2012). It has been argued that a dietary supply of *N*-acetyl neuraminic acid (sialic acid), a major constituent of acidic oligosaccharides in milk, may be essential for synthesis of brain gangliosides and the polysialic chains on neural cell adhesion molecules in preterm and rapidly growing infants (Wang and Brand-Miller 2003; Wang 2009). This argument is especially pertinent to mammals with altricial young, whose brains at birth may be at a less developed stage than early (e.g., 24 week) preterm human infants. We hypothesize that milk oligosaccharides and especially sialic acid-containing oligosaccharides are enriched in the milk of carnivoran species with altricial young in which a large proportion of brain and organ growth occurs postnatally.

A classic method for determining physical maturity at birth is to assess water content of lean tissues since this parameter gradually declines with development. When compared at birth across terrestrial (altricial) carnivorans, lean tissue water content ranges from 81.0% (domestic dog), 82.0% (domestic cat), 83.0% (mink) to 84.0% (American black bear) (Oftedal et al. 1993). Comparable values for precocial marine carnivores (Caniformia: Pinnipedia) are 71–74% ( $n = 4$  species) (Oftedal et al. 1996). The higher water content of lean tissue in terrestrial species is related to the less mature state of the neonates. However, little if any data are available for neonates of other carnivorans that would allow broad comparisons of oligosaccharide patterns.

An alternative approach is to examine, via mass assessment, postnatal development of the body, or specific organs, in relation to adult state. For example, brain mass at birth, expressed relative to adult mass, is considered a measure of the degree of neonatal maturity (Eisert et al. 2014). In order to maximize useful data, we examined total mass at birth of neonates, expressed as a percentage of maternal mass, with the assumption that a smaller neonate at birth will in general be less developed than larger neonates. Using species-specific data assembled by Oftedal and Gittleman (1989), the birth mass percentage of cheetahs (0.48) was considerably higher than American black bears (0.29), and this difference also holds at a familial level Felidae (cats)

(mean =  $1.60 \pm$  standard deviation (SD)  $0.80$ ,  $n = 13$ ) versus Ursidae (bears) (mean =  $0.30 \pm 0.08$  SD,  $n = 4$ ). Thus, these patterns are consistent with our predictions: altricial ursids have high OS:lactose ratios (10:1–52:1) compared to more precocial felids (1:1–1:2, see above).

Our data were sufficient to examine patterns among other Caniform and Feliform families. For example, three Caniform families known to have high OS:lactose ratios (Mephitidae, Mustelidae, Procyonidae) have relatively high but varied birth mass percentages (averaging  $2.67 \pm 0.45$ ,  $n = 2$ ;  $2.05 \pm 1.26$ ,  $n = 13$ ; and  $4.03 \pm 3.08$ ,  $n = 4$ ; respectively). Thus, their milks are higher in oligosaccharides than would be predicted from birth mass percentages. Therefore, further research is needed on both physiological maturity and postnatal development to determine correlations to milk oligosaccharide composition.

Unlike other Caniform species studied, lactose predominates over oligosaccharides in the milk of domestic dogs (Bubb et al. 1999). The neonatal dog is rather immature and does not open its eyes until around 10 days of age, but calculating the proportional size of puppies relative to the adult body weight is complicated by the extreme variation in size among different breeds of dogs. For example, puppies of giant breeds such as the English Mastiff may represent < 1% of adult BW, whereas puppies of toy breeds such as Papillon may be > 5% of adult BW at birth (Gropetti et al. 2017; Hawthorne et al. 2004; Scantlebury et al. 2000). Using Oftedal and Gittlemen's data for dogs (1989), a medium size dog has a neonatal weight representing 1.67% of maternal weight, aligning with our hypothesis that OS:lactose is lower in species having higher proportional mass at birth. However, the domestication of this species by humans may have influenced its lactational physiology, such that the concentration of oligosaccharides in milk may have become reduced relative to other Caniforms. To clarify this, future studies are needed to characterize the milk of non-domesticated canidae species such as wolf, coyote, red fox and bush dog.

### 15.4.2 *Artiodactyla*

The Artiodactyla represents another opportunity to seek biological correlates for type and quantity of oligosaccharide in colostrum and milk. First, Artiodactyls have much or somewhat lower oligosaccharide: lactose ratios than Carnivores: 1:1.6–1:22 (Artiodactyla) versus 52:1–1:1 (Carnivora) (Table 15.2). According to one hypothesis (above), higher oligosaccharide: lactose ratios may be associated with immaturity at birth. Based on body composition data of neonatal whitetail deer, cattle and horses, their percentage body water (% lean mass) is 79.4%, 75.7% and 74.6, respectively (Oftedal 1985). This is considerable lower than terrestrial Carnivores (81–84%), consistent with the observation that most or all Artiodactyls are precocial at birth and that their milks are lower in oligosaccharides than altricial Carnivores. The comparison of oligosaccharides: lactose ratios to birth mass, as an indicator of developmental state, is of limited value because of the tremendous interspecies variation in body mass. In one birth weight analysis, representing adult female masses of 4.5 kg

(dik-dik) through 316 kg (highveld eland) to 964 kg (giraffe), neonatal birth mass was equal to 14.9% (of maternal weight, dik-dik), 8.2% (eland) and 5.7% (giraffe) (Ofstedal 1985). It is doubtful that such great differences in relative mass at birth bear any relationship to developmental state, but rather reflect allometric effects. Thus, we would not expect relative birth masses to explain intra- or interordinal variation in milk oligosaccharides.

### ***15.4.3 Specific Milk Oligosaccharides: Are They Important for Artificial Feeding of Neonates?***

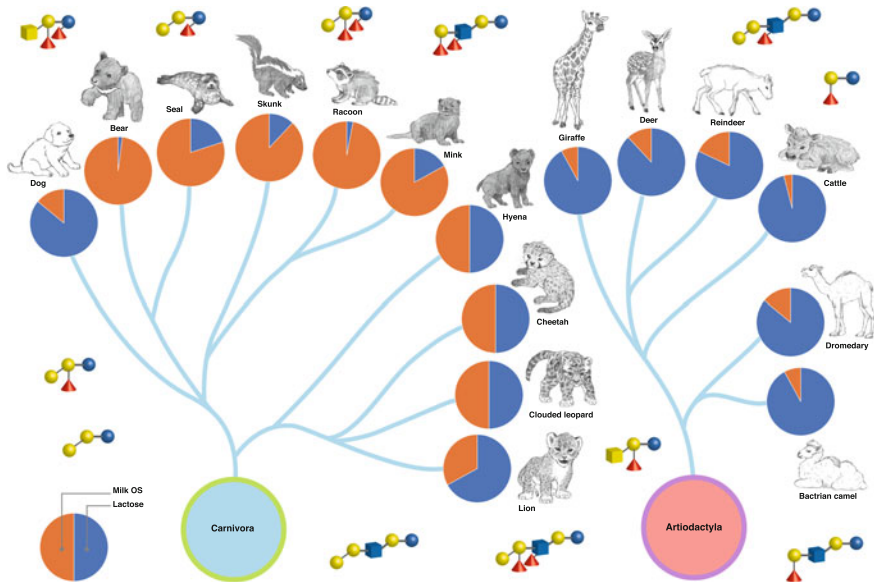
Despite the recognition that the aggregate amount of oligosaccharides in milk may exceed that of lactose, and the proliferation of information about the neutral and acidic oligosaccharide structures found among carnivoran milks, there is still little certainty about the physiological, nutritional, immunological or developmental necessity of these constituents for neonates (but see discussion of sialic acid, above). Studies are complicated by the need to be species-specific and the uncertainty about the extent to which findings are applicable to other, even closely related species. Although milk replacers continue to be produced by various manufacturers for dogs and cats, it is unlikely that these products match the oligosaccharide profiles of dog or cat milk, let alone the profiles of the milks of other zoo animals and wildlife for which they are employed. It is not known if this creates health or developmental problems. However, given that strong selective evolutionary pressures have apparently maintained the synthesis of a great diversity of oligosaccharide structures in carnivoran milks, especially in taxa with immature neonates, it seems likely that the oligosaccharides will prove to be important to neonatal development, both in general and at a taxon-specific level.

Among endangered species captive breeding programs, the artificial rearing of neonates occurs for a range of reasons, but particularly due to maternal ill-health, poor mothering/milk transfer and neonatal injury. For example, the cheetah is often reared on milk replacer formulated for domestic canine or feline requirements, but these requirements may poorly known and the formulas are based on bovine milk (Bell et al. 2010, 2012). But, the ratio of milk oligosaccharides to lactose and also the profile of the oligosaccharide in milk differ between these Carnivora species and the cow. Note that bovine mature milk contains only trace amount of oligosaccharides (Urashima et al. 2016), suggesting that some artificial oligosaccharides should be tested for use in milk replacers fed to carnivore species. The issue may be less pressing for milk-fed Artiodactyls, as cow's milk would be expected to be a good model.

## 15.5 Conclusion

The graphic image of the conclusion of this study is shown in Fig. 15.4. With the exception of the domestic dog, milk oligosaccharides predominate over lactose in the milk of Caniformia, including mink, striped skunk, raccoon, many bears and seals, whereas lactose is the dominant saccharide in the milk of some species of Feliformia, such as spotted hyena, African lion, clouded leopard and cheetah. The significant feature of the milk oligosaccharides of Carnivora is the presence of A ( $\text{GalNAc}\alpha 1-3(\text{Fuc}(\alpha 1-2)\text{Gal})$ ), B ( $\text{Gal}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}$ ) or H ( $\text{Fuc}\alpha 1-2\text{Gal}$ ) units as well as  $\alpha$ -Gal ( $\text{Gal}\alpha 1-3\text{Gal}$ ) unit attached to the core structures of lactose, lacto-*N*-neotetraose ( $\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3\text{Gal}\beta 1-4\text{Glc}$ ), lacto-*N*-neohexaose ( $\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3(\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-6)\text{Gal}\beta 1-4\text{Glc}$ ) or para lacto-*N*-neohexaose ( $\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3\text{Gal}\beta 1-4\text{Glc}$ ). The presence of A, B, H antigens, and  $\alpha$ -Gal vary depending on each species of Carnivora.

In contrast to the milk of Carnivora that of Artiodactyla including giraffe, sitatunga, deer and water buffalo contains lactose as the predominant saccharide, although small amounts of oligosaccharides are present as well. In most of these oligosaccharides, the core structure is lactose and they all contain isoglobotriose, but they are heterogeneous with respect to the presence of a few saccharides such as GM2 tetrasaccharide and globotriose.



**Fig. 15.4** Summary: Graphic representation of oligosaccharides in Carnivora and Artiodactyla. Pie graphs indicate relative proportions of lactose (blue) and oligosaccharides (orange); representative oligosaccharide structures are illustrated. (The illustration was done by MAKIKO GOTO)

## References

- Albrecht S, Lane JA, Marino K, Al Busadah KA, Carrington SD, Hickey RM, Rudd PM (2014) A comparative study of free oligosaccharides in the milk of domestic animals. *Br J Nutr* 111:1313–1328
- 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
- Bell KM, Rutherford SM, Cottam YH, Hendriks WH (2010) Evaluation of two milk replacers fed to hand-reared cheetah cubs (*Acinonyx jubatus*): Nutrient composition, apparent total tract digestibility, and comparison to maternal cheetah milk. *Zoo Biol* 29:1–15. <https://doi.org/10.1002/zoo.20344>
- Bell KM, Rutherford SM, Morton RH (2012) Growth rates and energy intake of hand-reared cheetah cubs (*Acinonyx jubatus*) in South Africa. *J Anim Physiol Anim Nutr* 96:182–190. <https://doi.org/10.1111/j.1439-0396.2011.01133.x>
- Bode L (2012) Human milk oligosaccharides: every baby needs a sugar mama. *Glycobiology* 22:1147–1162
- Bubb WA, Urashima T, Kohso K, Nakamura T, Arai I, Saito T (1999) (Occurrence of an unusual lactose sulfate in dog milk. *Carbohydr Res* 318:123–128
- Craft KM, Gaddy JA, Townsend SD (2018) Human milk oligosaccharides (HMOs) sensitize group B *Streptococcus* to clindamycin, erythromycin, gentamicin, and minocycline on a strain specific basis. *ACS Chem Biol* 13:2020–2026
- Duncan PI, Raymond F, Fuerholz A, Sprenger N (2009) Sialic acid utilization and synthesis in the neonatal rat revisited. *PLoS ONE* 4(12):e8241. <https://doi.org/10.1371/journal.pone.0008241>
- Eisert R, Potter CW, Oftedal OT (2014) Brain size in neonatal and adult Weddell seals: costs and consequences of having a large brain. *Marine Mammal Sci* 30(1):184–205. <https://doi.org/10.1111/mms.12033>
- Fukuda K, Yamamoto Y, Ganrorig K, Khuukhenbaatar J, Senda A, Saito T, Urashima T (2010) Chemical characterization of the oligosaccharides in Bacterian camel (*Camelus bactrianus*) milk and colostrum. *J Dairy Sci* 93:5572–5587
- Gal B, Ruano MJ, Puente R, Garcia-Pardo LA, Rueda R, Gil A, Hueso P (1997) Developmental changes in UDPN-acetylglucosamine 2-epimerase activity of rat and guinea-pig liver. *Comp Biochem Physiol B: Biochem Mol Biol* 118:13–15
- Ganzorig K, Asakawa T, Sasaki M, Saito T, Suzuki I, Fukuda K, Urashima T (2018) Identification of sialyl oligosaccharides including an oligosaccharide nucleotide in colostrum of an addax (*Addax nasomaculatus*) (Subfamily Antelopinae). *Anim Sci J* 89:167–175
- Griffiths K, Hou R, Wang R, Zhang Z, Zhang L, Zhang T, Watson DG, Burchmore RJS, Loeffler KL, Kennedy MW (2015) Prolonged transition time between colostrum and mature milk in a bear, the giant panda, *Ailuropoda melanoleuca*. *R Soc Open Sci* 2:150395
- Groppetti D, Pecile A, Palestini C, Marelli SP, Boracchi P (2017) A national census of birth weight in purebred dogs in Italy. *Animal* 7:43–63
- Hawthorne AJ, Booles D, Nugent PA, Gettinby G, Wilkinson J (2004) Body-weight changes during growth in puppies of different breeds. *J Nutr* 134:2027S–2030S
- He Y, Liu S, Kling DE, Leone S, Lawlor NT, Huang Y, Feinberg SB, Hill DR, Newburg DS (2016) The human milk oligosaccharide 2'-fucosyllactose modulates CD14 expression in human enterocytes, thereby attenuating LPS-induced inflammation. *Gut* 65:33–46
- Jenness GA, Regehr EA, Sloan RE (1964) Comparative studies of milks. II. Dialyzable carbohydrates. *Comp Biochem Physiol* 13:339–352
- Lin AE, Autran CA, Szyszka A, Escajadillo T, Huang M, Godula K, Prudden AR, Boons GJ, Lewis AL, Donovan KS, Nizet V, Bode L (2017) Human milk oligosaccharides inhibit growth of group B *streptococcus*. *J Biol Chem* 292:11243–11249
- Messer M, Urashima T (2002) Evolution of milk oligosaccharides and lactose. *Trends Glycosci Glycotechnol* 14:153–176

- Mineguchi Y, Miyoshi M, Taufik E, Kawamura A, Asakawa T, Suzuki I, Souma K, Okubo M, Saito T, Fukuda K, Asakuma S, Urashima T (2018) Chemical characterization of the milk oligosaccharides of some Artiodactyla species including giraffe (*Giraffa camelopardalis*), sitatunga (*Tragelaphus spekii*), deer (*Cervus nippon yesoensis*) and water buffalo (*Bubalus bubalis*). *Glycoconj J* 35:561–574
- Nakamura T, Urashima T, Mizukami T, Fukushima M, Arai I, Senshu T, Imazu K, Nakao T, Saito T, Ye Z, Zuo H, Wu K (2003a) Composition and oligosaccharides of a milk sample of the giant panda, *Ailuropoda melanoleuca*. *Comp Biochem Physiol B* 135:439–448
- Nakamura T, Kawase H, Kimura K, Watanabe Y, Ohtani M, Arai I, Urashima T (2003b) Changes in bovine colostrum and milk sialyloligosaccharides during early lactation. *J Dairy Sci* 86:1315–1320
- Oftedal OT (1985) Pregnancy and lactation. In: Hudson RJ, White RG (eds) *The bioenergetics of wild herbivores*. CRC Press, Boca Raton, FL, pp 215–238
- Oftedal OT, Gittleman JG (1989) Patterns of energy output during reproduction in carnivores. In: Gittleman JG (ed) *Carnivore behavior, ecology and evolution*. Cornell University Press, Ithaca, NY, pp 375–378
- Oftedal OT, Alt GL, Widdowson EM, Jakubasz MR (1993) Nutrition and growth of suckling black bears (*Ursus americanus*) during their mothers' winter fast. *Brit J Nutr* 70:59–79
- Oftedal OT, Bowen WD, Boness DJ (1996) Lactation performance and nutrient deposition in pups of the harp seal, *Phoca groenlandica*, on ice floes off southeast Labrador. *Physiol Zool* 69:635–657
- Rostami SM, Bebet T, Spears J, Reynolds A, Satyaraj E, Sprenger N, Austin S (2014) Milk oligosaccharides over time of lactation from different dog breeds. *PLoS ONE* 9:e99824
- Sasaki M, Nakamura T, Hirayama K, Fukuda K, Saito T, Urashima T, Asakuma S (2016) Characterization of two novel sialyl N-acetyllactosaminyl nucleotides separated from ovine colostrum. *Glycoconj J* 33:789–796
- Scantlebury M, Butterwick R, Speakman JR (2000) Energetics of lactation in domestic dog (*Canis familiaris*) breeds of two sizes. *Comp Biochem Physiol A* 127:197–210
- 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
- Sprenger N, Duncan PI (2012) Sialic acid utilization. *Adv Nutr* 3(3): 392S–397S. <https://doi.org/10.3945/an.111.001479>
- Tao N, Wu S, Kim J, Joo An H, Hinde K, Power M, Gagneux P, German JB, Lebrilla CB (2011) Evolutionary glycomics: characterization of milk oligosaccharides in primates. *J Proteome Res* 10:1548–1557. <https://doi.org/10.1021/pr1009367>
- 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
- Taufik E, Ganzorig K, Nansalma 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
- 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
- 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 A* 152:158–161
- Urashima T, Saito T, Ohmisy K, Shimazaki K (1991) Structural determination of three neutral oligosaccharides in bovine (Holstein-Friesian) colostrum, including the novel trisaccharide; GalNAc $\alpha$ 1-3Gal $\beta$ 1-4Glc. *Biochim Biophys Acta* 1073:225–229
- Urashima T, Kusaka Y, Nakamura T, Saito T, Maeda N, Messer M (1997) Chemical characterization of milk oligosaccharides of the brown bear, *Ursus arctos yesoensis*. *Biochim Biophys Acta* 1334:247–255

- Urashima T, Sumiyoshi W, Nakamura T, Arai I, Saito T, Komatsu T, Tsubota T (1999a) Chemical characterization of milk oligosaccharides of the Japanese black bear, *Ursus thibetanus japonicus*. *Biochim Biophys Acta* 1472:290–306
- Urashima T, Yamamoto M, Nakamura T, Arai I, Saito T, Namiki M, Yamaoka K, Kawahara K (1999b) Chemical characterisation of the oligosaccharides in a sample of milk of a white-nosed coati, *Nasua narica* (Procyonidae: Carnivora). *Comp Biochem Physiol A* 123:187–193
- 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, Arita M, Yoshida M, Nakamura T, Arai I, Saito T, Arnould JPY, Kovacs KM, Lydersen C (2001) Chemical characterization of the oligosaccharides in hooded seal. (*Cystophora cristata*) and Australian fur seal (*Arctocephalus pusillus doriferus*) milk. *Comp Biochem Physiol B* 128:307–323
- 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 B* 132:611–624
- Urashima T, Nakamura T, Yamaguchi K, Munakata J, Arai I, Saito T, Lydersen C, Kovacs KM (2003) Chemical characterization of the oligosaccharides in milk of high Arctic harbour seal (*Phoca vitulina vitulina*). *Comp Biochem Physiol A* 135:549–563
- Urashima T, Nakamura T, Teramoto K, Arai I, Saito T, Komatsu T, Tsubota T (2004) Chemical characterization of sialyl oligosaccharides in milk of the Japanese black bear, *Ursus thibetanus japonicus*. *Comp Biochem Physiol B* 139:587–595
- 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 A* 142:461–471
- 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 B* 146:153–159
- Urashima T, Fukuda K, Messer M (2012) Evolution of milk oligosaccharides and lactose: a hypothesis. *Animal* 6(3):369–374
- Urashima T, Messer M, Oftedal OT (2014) Comparative biochemistry and evolution of milk oligosaccharides of monotremes, marsupials, and eutherians. In: Pontarotti P (ed) *Evolutionary biology: genome evolution, speciation, coevolution and origin of life*. Springer, Switzerland, pp 3–33
- Urashima T, Messer M, Oftedal OT (2016) Oligosaccharides in the milk of other mammals. In: Mi M, Ma M, Bode L (eds) *Prebiotics and probiotics in human milk*. Academic Press, Amsterdam, pp 45–139
- Urashima T, Messer M (2017) Evolution of milk oligosaccharides and their function in monotremes and marsupials. In: Pontarotti P (ed) *Evolutionary biology: self/nonself evolution, species and complex traits evolution, methods and concepts*. Springer, Switzerland, pp 237–256
- Urashima T, Hirabayashi J, Sato S, Kobata A (2018a) Human milk oligosaccharides as essential tools for basic and application studies on galectins. *Trends Glycosci Glycotechnol* 30: SE51–65
- Urashima T, Yamaguchi E, Ohshima T, Fukuda K, Saito T (2018b) Chemical structures of oligosaccharides in milk of the raccoon (*Procyon lotor*). *Glycoconj J* 35: 275–286
- Urashima T, Umewaki M, Taufik E, Ohshima T, Fukuda K, Saito T, Whitehouse-Tedd W, Budd JA, Oftedal OT (2020) Chemical structures of oligosaccharides in milk of American black bear (*Ursus americanus americanus*) and cheetah (*Acinonyx jubatus*). *Glycoconj J* 37: 57–76
- Wang B, Brand-Miller J, McNeil Y, McVeagh P (1998) Sialic acid concentration of brain gangliosides: variation among eight mammalian species. *Comp Biochem Physiol* 119A:435–439
- Wang B, Brand-Miller J (2003) The role and potential of sialic acid in human nutrition. *Eur J Clin Nutr* 57:1351–1369. <https://doi.org/10.1038/sj.ejcn.1601704>

- Wang B (2009) Sialic acid is an essential nutrient for brain development and cognition. *Annu Rev Nutr* 29:177–222. <https://doi.org/10.1146/annurev.nutr.28.061807.155515>
- Yu ZT, Nanthakumar NN, Newbuug DS (2016) The human milk oligosaccharides 2'-fucosyllactose quenches *Campylobacter jejuni*-induced inflammation in human epithelial cells HEP-2 and HT-29 and in mouse intestinal mucosa. *J Nutr* 146:1980–1990



# Chapter 16

## Making Sense of Noise



Shu-Ting You and Jun-Yi Leu

**Abstract** Noise is the heterogeneity in transcript or protein levels existing in an isogenic population under the same growth condition. Expression noise is unavoidable in living cells. To ensure accurate execution of cellular functions, cells have developed several mechanisms to reduce noise. However, noise can also be utilized to facilitate different levels of regulation. With advances in single-cell analyses and genomics tools, we now know that noise influences almost every aspect of life. Here, we first present the regulatory systems underlying noise, from general principles to specific molecular mechanisms. Additionally, we discuss an experimental evolution approach for finding new mechanisms involved in noise regulation. Next, we review the evolutionary implications of noise from immediate benefits to long-term population survival. We explore the interactions between noise and mutations. Finally, we briefly discuss how noise can impact inter-population interactions, representing a possible link between micro-scale and macro-scale biological phenomena.

### 16.1 From Phenotypic Variation to Noise

Cells of the same genotype often exhibit a variety of phenotypes depending on their living conditions. Even when existing in the same environment, isogenic cells of the same cell cycle stage and age can still present phenotypic variation. Such individuality in an isogenic population (non-genetic variation) was first recognized in 1945 from bacteriophage infectivity (Delbruck 1945). Bacteriophages multiply within host cells before lysing the hosts to infect new ones. Although host cells were all initially infected by a single bacteriophage, copies of bacteriophage liberated from

---

S.-T. You (✉)

Department of Molecular Genetics, Weizmann Institute of Science, 234 Herzl Street, POB 26, Rehovot 7610001, Israel  
e-mail: [biosupernatant@gmail.com](mailto:biosupernatant@gmail.com)

J.-Y. Leu

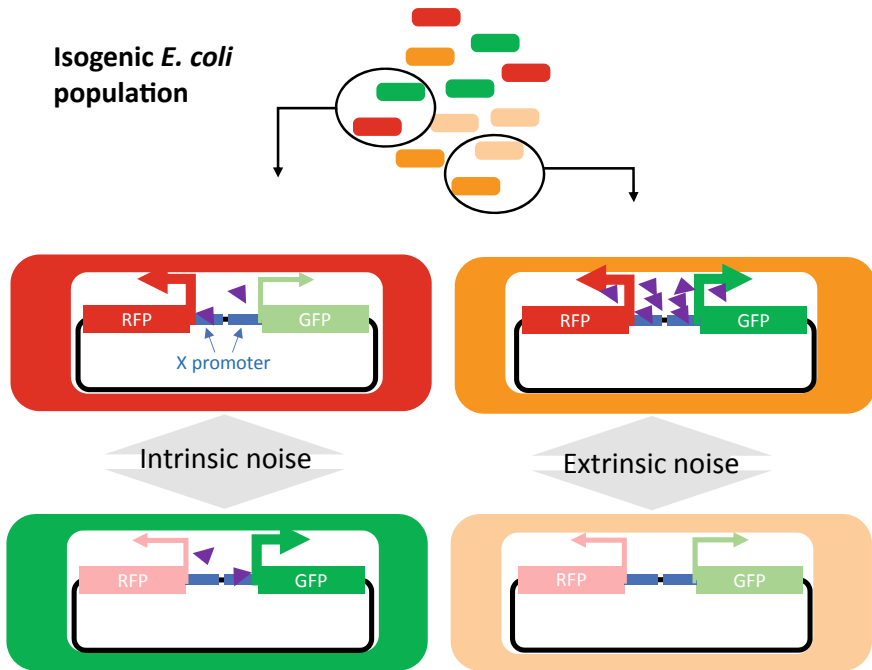
Institute of Molecular Biology, Academia Sinica, 128 Sec. 2, Academia Road, Nankang, Taipei 115, Taiwan  
e-mail: [jleu@imb.sinica.edu.tw](mailto:jleu@imb.sinica.edu.tw)

different host cells varied more than 20-fold (Delbruck 1945). With the development of single-cell technologies in recent years, non-genetic variation has become ubiquitously observed, from microbes to mammalian cells and from normal to cancer cells (Avery 2006; Balazsi et al. 2011; Brock et al. 2009). Such individuality can arise from different molecular processes, including DNA and histone modifications, transcription, mRNA splicing, circular RNA production, translation, and protein folding (Chalancon et al. 2012; Hu and Zhou 2018; Itakura et al. 2020; Kim and Jacobs-Wagner 2018; Mikl et al. 2019). Furthermore, some non-genetic variation can persist for more than one generation, indicative of a cross-generational effect. In this article, we discuss the general molecular mechanisms and biological impacts of non-genetic variation. We use the term “noise” to represent the heterogeneity at either mRNA or protein level that is the molecular outcome of non-genetic variation.

Noise is unavoidable in living cells. Elowitz and his colleagues elegantly showed that non-correlated signals between two fluorescent proteins could be observed in a single cell, even though the two reporter genes were controlled by identical promoters and located the same distance from the common replication origin (Fig. 16.1) (Elowitz et al. 2002). They reasoned that many steps in the process of protein production are initiated through binding reactions, such as DNA binding of RNA polymerases and mRNA binding of ribosomes. Factor binding, like a chemical reaction, has an inherent probability of manifesting cell-to-cell heterogeneity. Noise resulting from such events is called intrinsic noise, as it can be revealed in a single cell. Another inevitable noise occurs during cell division. When cellular material (e.g., RNAs and proteins) is redistributed from mother to daughter cells, much of it is subjected to probabilistic segregation, leading to variation between individual cells (Golding et al. 2005; Huh and Paulsson 2011; Kinkhabwala et al. 2014). If unevenly partitioned proteins belong to basal machineries (e.g., RNA polymerase and ribosome) or upstream regulators, the resulting variation can be further propagated, inducing cell-to-cell heterogeneity in many downstream proteins, which is termed extrinsic noise. The aforementioned two-fluorescent-protein system can also reveal extrinsic noise, as the differences in intensity of fluorescence signal between cells are greater than for within-cell differences (Fig. 16.1) (Elowitz et al. 2002). Thus, the noise of a given gene is composed of intrinsic and extrinsic elements, and it can further influence other genes through the interaction network (Pedraza and van Oudenaarden 2005; Raser and O’Shea 2004).

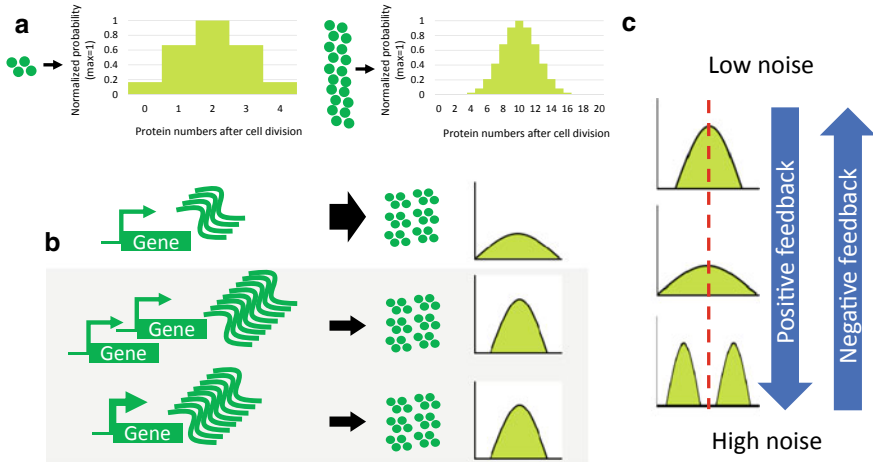
## 16.2 General Strategies for Noise Regulation

Cells continually process genetic information and produce proteins to maintain normal physiology. The levels and timing of protein production usually need to be precisely controlled, with excessive noise potentially being harmful. Cells deploy several general strategies to reduce noise. One common strategy is to increase the number of “participants” in cellular events, including gene copy numbers, transcript levels, or numbers of key regulatory molecules. The stochastic nature of noise



**Fig. 16.1** Visualizing the noise in an isogenic population growing in the same environment. Elowitz and his colleagues (2002) generated an *E. coli* strain expressing two different fluorescent proteins (RFP and GFP) to demonstrate noise within an isogenic population. The fluorescent protein-encoding genes were driven by the same type of promoters. Intrinsic noise was revealed by individual cells exhibiting different ratios of RFP/GFP intensity. Intrinsic noise could result from stochastic binding of the transcription factor to the promoter. The stochasticity is inherent in any binding events between molecules. Therefore, intrinsic noise is unavoidable at molecular levels, even when the numbers of regulatory molecules are homogenous. Alternatively, noise due to different amounts (or activities) of regulatory molecules is termed extrinsic noise, as revealed by the difference in fluorescence intensity between cells being greater than the difference within cells

has a profound effect on events involving low participant numbers (Maheshri and O’Shea 2007; Raser and O’Shea 2005). For example, when a cytosolic protein X is randomly distributed in dividing cells, the proportional difference in levels of X between the two daughter cells typically increases if the copy number of X is reduced (Fig. 16.2a). Indeed, an anti-correlation between mean expression levels and noise has been reported, with the relationship being more evident for low-expression genes (Bar-Even et al. 2006; Newman et al. 2006; Wu et al. 2017). Consequently, cells may display increased duplicate gene copies or transcript levels but lower translation rates (so the final protein abundance remains the same) to control noise (Fig. 16.2b). In yeast, essential proteins or complex-forming proteins often exhibit higher transcription rates and lower translation rates to minimize their expression noise (Fraser et al. 2004). Moreover, cells in the G2 phase of the cell cycle exhibit lower noise than those in G1 since G2 cells have double the number of gene copies (Keren et al. 2015).



**Fig. 16.2** Noise regulation. **a** Noise derived from random segregation at cell division is profound for proteins of low abundance. After cell division, cells with extreme molecule numbers of the protein are more frequent for the lowly abundant protein (left panel) compared to cells for the highly abundant one (right panel). **b** Increasing gene copies (middle panel) or mRNA levels (bottom panel) reduces protein production noise. Copy numbers of genes and transcripts are relatively low compared to the number of proteins, so they are more sensitive to fluctuation. When the copy numbers of genes or mRNA increase, the corresponding translation efficiency is reduced to maintain the same protein level. **c** Noise is up-regulated by positive feedback and down-regulated by negative feedback without altering mean expression. One extreme example of high noise is a bivalent population

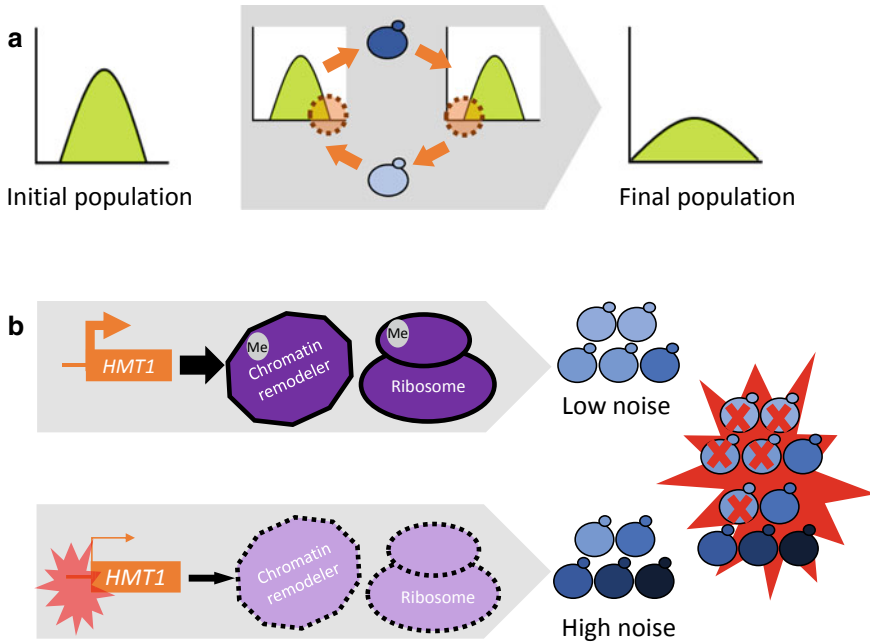
Noise reduction can also be achieved through negative feedback networks, whereby gene expression is suppressed by the encoded transcripts or proteins or by downstream products. Such negative feedback modules are found in many regulatory networks within cells, enabling expression to be maintained while eliminating unwanted fluctuations (Alon 2007). In addition to negative feedback, other network structures can also influence the level of noise. Autoregulation through positive feedback increases noise via a snowball effect. For instance, a transient increment in expression of a given gene can be sustained because the increment stimulates self-expression and intensifies gene activation. Similar noise amplification is apparent in double-negative feedback loops, whereby two genes can suppress each other, often resulting in a switch between on and off states for the genes. Accordingly, extreme differences between cells can be revealed when gene expression is monitored within a cell population (Fig. 16.2c) (Alon 2007; Ferrell 2002). In the cell, these network structures are integrated to regulate the level and noise of gene expression (Chalancón et al. 2012). However, if we look at individual genes, finding that a gene is regulated by both negative and auto-regulatory positive feedback is not rare. The integrative nature of biological networks imposes a host of challenges for deciphering noise regulation. To better understand how cells control noise, researchers need to explore molecular mechanisms in detail and investigate how they crosstalk with each other.

### 16.3 Molecular Mechanisms Involved in Noise Regulation

Various molecular mechanisms have been observed to influence noise at different levels, including gene promoter architectures, nucleosome occupancy, histone modification, cellular compartmentalization, microRNAs, protein chaperoning, and signal transduction (Arias and Hayward 2006; Battich et al. 2015; Hsieh et al. 2013; Sanchez et al. 2013; Schmiedel et al. 2015; Weinberger et al. 2012; Wu et al. 2017). These mechanisms may work independently or in concert to regulate the noise of individual genes (Faure et al. 2017). For example, some lowly expressed essential genes that should suffer from high intrinsic noise have been shown to utilize histone modifications to acquire low noise status (Wu et al. 2017). Similarly, microRNAs can preferentially target lowly expressed genes to reduce their protein expression noise (Schmiedel et al. 2015). Cellular compartmentalization in the nucleus or cytoplasm has been suggested to function as a global noise filter in mammalian cells, but it may also work together with other mechanisms to fine-tune gene-specific noise (Battich et al. 2015).

Theoretically, noise regulation can be condition-specific or tissue-specific (e.g., by tissue-specific histone modifications or microRNA expression). In yeast, Hsp90 protein chaperoning has been shown to enhance non-genetic variation upon encountering heat stress. It is speculated that this enhanced noise may increase the likelihood of population survival under stress conditions (Hsieh et al. 2013). Under aerobic conditions, *Escherichia coli* cells also present increased noise of a signal transduction pathway, which permits tolerance to rapid environmental change (Carey et al. 2018). Condition-specific or tissue-specific noise regulation allows organisms to prevent the deleterious impacts of noise under normal conditions and to enhance population complexity when required. Nonetheless, it remains unclear how commonly cells exploit these types of regulation in natural environments or during development.

Identifying novel molecular mechanisms of noise regulation is always challenging. Harnessing the power of experimental evolution, researchers can delineate the genetic basis of complex traits (McDonald 2019). To uncover pathways involved in noise regulation, previously we imposed a selection scheme to favor yeast cells that alter their levels of reporter proteins both markedly and frequently under a normal growth condition (Fig. 16.3a). Our rationale for that approach is that high noise is usually inhibited when cells are growing under normal conditions. However, if the cells harbor mutations that disrupt noise regulators, then they are more likely to produce extremely low or high levels of reporter proteins, which would be enriched by our selection regime. By analyzing the mutations existing in our evolved populations, we identified a methyltransferase that coordinates noise in response to environmental conditions (Fig. 16.3b) (You et al. 2019). This gene suppresses the noise of multiple pathways when environments are benign. However, its expression is down-regulated by various environmental stresses, resulting in increased noise. Since a population with high heterogeneity survives better under stress, this regulatory strategy for modulating noise is likely a product of natural selection. The methyltransferase we identified is a hub protein in yeast protein networks. Its number of interacting



**Fig. 16.3** A noise regulator coordinating internal and external environments. **a** An alternating selection scheme enriches cells harboring mutations that interfere with noise reduction. Cells with one extreme level of the reporter protein are selected, expanded, and serve as a parental population for selecting cells with the other extreme. When the distribution of the reporter protein becomes consistently flat, the final population is analyzed to identify the noise-enhancing mutations. **b** Hmt1 methylates target proteins to regulate the functions of several protein complexes, including chromatin remodelers and ribosomes. Hmt1 levels under normal conditions are sufficiently high to modulate downstream proteins, resulting in homogeneous gene expression among cells (upper panel). When the population is challenged by environmental stresses, *Hmt1* expression is down-regulated. Consequently, the target protein complexes are compromised and gene expression becomes noisier (lower panel). Accordingly, heterogeneous gene expression results in phenotypic variations between individual cells and further promotes the likelihood of population survival. Me, methylation

partners ranks in the top 99th percentile of the yeast interactome (Stark et al. 2006) explaining why it has a general effect on many pathways. We also demonstrated that the noise regulation by this enzyme is conserved between two yeast species that have diverged for 500 million years. Furthermore, both the enzyme and its downstream targets (which also suppress noise) have orthologs in human cells, suggesting that they may also regulate noise in multicellular organisms. Our identification of a new noise regulatory pathway demonstrates that experimental evolution is a promising approach for exploring such noise modulatory mechanisms. By applying different reporter genes and selection conditions, researchers are likely to uncover other such molecular mechanisms underlying noise regulation.

## 16.4 The Biological Impacts of Noise

It appears that noise levels of individual genes are shaped by natural selection. For example, genes in the same pathways have similar noise levels, suggesting that they have evolved pathway-specific regulation (Stewart-Ornstein et al. 2012). Since fluctuations in basal machineries (e.g., RNA polymerase) are more likely to impact cellular fitness, genes encoding such basal machineries often exhibit lower noise than environment-responsive genes (e.g., heat shock protein) (Bar-Even et al. 2006; Newman et al. 2006). Moreover, genes involved in multiple pathways also exhibit low noise, perhaps due to simultaneous constraints from different pathways (Barroso et al. 2018).

Cells have developed many noise-constraining mechanisms to ensure accurate cellular processes. However, they also actively exploit noise under certain conditions (Table 16.1). The risk-spreading bet-hedging strategy is one example of noise exploitation commonly observed in microbes (de Jong et al. 2011; Simons 2011). Unicellular organisms constantly encounter environmental fluctuations. Although cells have evolved sophisticated environment-responsive pathways, they may still suffer under novel or abrupt environmental stresses. By utilizing the stochastic nature of protein or RNA production, even a clonal population can generate heterogeneity to manifest some extreme phenotypes (e.g., proliferating or quiescent) that can enhance the likelihood of population survival upon experiencing unexpected stresses (Balaban et al. 2004; Levy et al. 2012). This strategy allows cells to thrive in unpredictable environments without having to maintain overly complex environment-responsive

**Table 16.1** Benefits from noise

	Benefit	Applied strategy	Literature
Unicellular organisms in response to environments	Survival	Risk spreading	Balaban et al. (2004), Levy et al. (2012)
	Adaptation	Exploration through multiple trajectories	Freddolino et al. (2018)
		Increasing sensitivity	Paulsson et al. (2000), Wiesenfeld and Moss (1995)
Multicellular organisms in development	Energy-saving	Free from complex regulatory networks to initiate development	Losick and Desplan (2008)
	Sustainability	Balanced noise at different regulatory layers	Phillips et al. (2016)
Organism evolution and cancer development	Survival	Risk spreading	Brock et al. (2009)
	Mutagenicity	Oncogene activation	Cejas and Long (2020), Faure et al. (2017)
		Increased expression of mutators	Liu et al. (2019), Uphoff et al. (2016)

systems (which can be very costly). Even if environmental changes are not drastic, noise can facilitate adaptation to them since the co-existence of multiple cellular states in a population enables the cells to adopt the best adaptive solution from multiple starting points (Freddolino et al. 2018). When it acts upon individual signaling pathways, noise can allow cells to respond more sensitively. Firstly, noise boosts the under-threshold signals above the detection threshold, which is thought to operate in neuron-sensing systems. Secondly, when the time of noise generation is shorter than the half-life of responding products, the levels of product can be amplified without significantly dampening their subsequent effects (Paulsson et al. 2000; Wiesenfeld and Moss 1995).

In multicellular organisms, noise is also widely manipulated to initiate cell differentiation and stabilize cell commitment (Balazsi et al. 2011; Ferrell 2002; Hansen et al. 2018). Together with auto-regulatory positive or mutual negative feedback, noise can be used to switch a cell from an indecisive state to a deterministic one. In the case of photoreceptor development in fruit flies, the first wave of cell fate determination is initiated when randomly fluctuating signals exceed a threshold to trigger feedback. Once the fate of a cell is stochastically determined, neighboring cell fate is determined through cell–cell interactions, leading to a program-like outcome (Wernet et al. 2015). This energy-saving design allows fruit flies to develop sophisticated eye patterns without maintaining a complex regulatory circuit that is used only once in a lifetime (Losick and Desplan 2008). Interestingly, different layers of noise can counteract each other, leading to a robust outcome at the population level. Such a stochastic model has been invoked to explain the observed differentiation process of neuronal cells. Although several key regulators exhibit low abundance during cell division and present unequal distributions, this noise is absorbed by another layer of noise (i.e., noise in the regulatory network between these key molecules), so a robust population structure is maintained (Phillips et al. 2016).

In addition to developmental processes, noise is thought to influence the evolutionary trajectory of cancers. The continuously changing cell states in cancer development pose medical difficulties for cancer treatments since they provide a large number of variants from which cells with survival advantages can be selected (Brock et al. 2009). Like genome instability, elevated noise allows cancer cells to maintain high heterogeneity (Cohen et al. 2008). Rhabdoid cancer is one of the cancers driven by inactivation mutations in the SWI/SNF chromatin remodeler (Roberts et al. 2002), a protein complex that inhibits general expression noise (Raser and O’Shea 2004). Inactivation of the SWI/SNF complex also accelerates the tumorigenesis of aggressive cancers (Roberts et al. 2002). Similarly, super-enhancer elements are enriched at putative oncogenes for brain tumors and it is known that genes with super-enhancer elements tend to exhibit high noise (Cejas and Long 2020; Faure et al. 2017). Thus, phenotypic variation derived from noise might itself accelerate the process of tumorigenesis or work together with genome instability to exacerbate the overall effect. We discuss topics related to interactions between non-genetic and genetic variation in the next section.



## 16.5 Interactions Between Non-genetic and Genetic Variation

Although the effects of non-genetic and genetic variation persist for different time-scales, these two types of variation often crosstalk with each other to further influence cells and their evolution (Yona et al. 2015). In a small population, evolutionary rescue (meaning rescue from devastating environmental conditions by existing genetic variation) is often inaccessible when environments change swiftly. However, the presence of noise enables certain cells to survive and buys time for the whole population to acquire beneficial mutations (O’Dea et al. 2016). Interestingly, high noise is often associated with high mutation rates (Makova and Hardison 2015; Sanchez et al. 2013; Weinberger et al. 2012). Since the effects of *de novo* mutations are unknown (though they tend to be detrimental), the high noise of a given gene may not only alleviate harm by preserving cells that express low levels of that gene but also exploit any benefits from cells that highly express it. A few studies have shown that noise can even facilitate the generation of genetic variation by facilitating fluctuations in protein abundance of the key regulators involved in DNA replication, recombination or repair (Liu et al. 2019; Uphoff et al. 2016). Overall, noise can facilitate the retention of beneficial mutations, firstly by enabling the persistence of cells under stress, secondly by balancing mutational effects with population fitness and, finally, by generating further mutations.

Incomplete penetrance of mutational effects is a phenomenon by which the same mutation can differentially affect the fate of individual organisms in an isogenic population (Eldar et al. 2009). Incomplete penetrance has been observed in various biological systems, including human diseases (Gruber and Bogunovic 2020; Taubner et al. 2018). However, in most cases, the underlying molecular mechanisms remain elusive. Recent studies indicate that the interactions between genetic and non-genetic variation represent one of the possible mechanisms for incomplete penetrance (Burga et al. 2011; Raj et al. 2010). When mutations occur in a biological pathway, their deleterious effects can be masked by the existence of redundant genes (or pathways) or other buffering machinery (such as Hsp90) (Siegal and Leu 2014). If the expression of buffering genes is noisy, mutational effects will be differentially manifested depending on levels of the buffering genes. As discussed previously, certain types of noise are subjected to condition-specific or tissue-specific regulation, further indicating that the specific influences of environments or tissues in some cases of incomplete penetrance can also be explained by the noise–mutation interaction model (Cooper et al. 2013).

## 16.6 What Is Next?

Since Delbruck’s discovery of heterogeneity in bacteriophage infection (Delbruck 1945), our knowledge about biological noise has grown tremendously and noise has

been shown to influence almost every aspect of life. It is about time that we get back to Delbruck's original observation and rethink how noise impacts the interaction dynamics between different organisms. Connecting micro-scale phenomena (stochasticity at the molecular level) to macro-scale interspecific dynamics is challenging and interpreting respective data may be exceedingly complicated. A recent study established a multi-scale model to investigate bacteria–phage interaction. The results illustrated how optimal bacterial defense strategies are shaped in the presence of noise (Ruess et al. 2019). Phenotypic variation affecting the strength of interspecific interactions can further influence the fate of interacting populations. A consumer–resource ecological model predicts that high phenotypic variation weakens interactions and alleviates the effects of single consumers on single resources, subsequently stabilizing species co-existence (Gibert and Brassil 2014). In that case, noise could impact the stability of food webs, but that scenario also reveals how individual variation can contribute to the persistence of interacting populations. It will be interesting to further explore the role of noise in ecology, especially in microbe–microbe or microbe–host interaction systems that can be tested experimentally.

**Acknowledgements** Jun-Yi Leu was supported by Academia Sinica of Taiwan (grant no. AS-IA-105-L01 and AS-TP-107-ML06) and the Taiwan Ministry of Science and Technology (MOST108-2321-B-001-001). We thank John O'Brien for manuscript editing.

## References

- Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet* 8:450–461
- Arias AM, Hayward P (2006) Filtering transcriptional noise during development: concepts and mechanisms. *Nat Rev Genet* 7:34–44
- Avery SV (2006) Microbial cell individuality and the underlying sources of heterogeneity. *Nat Rev Microbiol* 4:577–587
- Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S (2004) Bacterial persistence as a phenotypic switch. *Science* 305:1622–1625
- Balazsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. *Cell* 144:910–925
- Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* 38:636–643
- Barroso GV, Puzovic N, Dutheil JY (2018) The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics* 208:173–189
- Battich N, Stoeger T, Pelkmans L (2015) Control of transcript variability in single mammalian cells. *Cell* 163:1596–1610
- Brock A, Chang H, Huang S (2009) Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours. *Nat Rev Genet* 10:336–342
- Burga A, Casanueva MO, Lehner B (2011) Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature* 480:250–253
- Carey JN, Mettert EL, Roggiani M, Myers KS, Kiley PJ, Goulian M (2018) Regulated stochasticity in a bacterial signaling network permits tolerance to a rapid environmental change. *Cell* 175:1989–1990
- Cejas P, Long HW (2020) Principles and methods of integrative chromatin analysis in primary tissues and tumors. *Biochim Biophys Acta Rev Cancer* 1873:188333

- Chalancon G, Ravarani CNJ, Balaji S, Martinez-Arias A, Aravind L, Jothi R, Babu MM (2012) Interplay between gene expression noise and regulatory network architecture. *Trends Genet* 28:221–232
- Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, Sigal A, Milo R, Cohen-Saidon C, Liron Y, Kam Z et al (2008) Dynamic proteomics of individual cancer cells in response to a drug. *Science* 322:1511–1516
- Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* 132:1077–1130
- de Jong IG, Haccou P, Kuipers OP (2011) Bet hedging or not? A guide to proper classification of microbial survival strategies. *BioEssays* 33:215–223
- Delbruck M (1945) The burst size distribution in the growth of bacterial viruses (bacteriophages). *J Bacteriol* 50:131–135
- Eldar A, Chary VK, Xenopoulos P, Fontes ME, Loson OC, Dworkin J, Piggot PJ, Elowitz MB (2009) Partial penetrance facilitates developmental evolution in bacteria. *Nature* 460:510–514
- Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186
- Faure AJ, Schmiedel JM, Lehner B (2017) Systematic analysis of the determinants of gene expression noise in embryonic stem cells. *Cell Syst* 5:471–484
- Ferrell JE (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr Opin Cell Biol* 14:140–148
- Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. *PLoS Biol* 2:e137
- Freddolino PL, Yang J, Momen-Roknabadi A, Tavazoie S (2018) Stochastic tuning of gene expression enables cellular adaptation in the absence of pre-existing regulatory circuitry. *eLife* 7:e31867
- Gibert JP, Brassil CE (2014) Individual phenotypic variation reduces interaction strengths in a consumer-resource system. *Ecol Evol* 4:3703–3713
- Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123:1025–1036
- Gruber C, Bogunovic D (2020) Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. *Hum Genet* 139:745–757
- Hansen MMK, Desai RV, Simpson ML, Weinberger LS (2018) Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability. *Cell Syst* 7(384–397):e386
- Hsieh YY, Hung PH, Leu JY (2013) Hsp90 regulates nongenetic variation in response to environmental stress. *Mol Cell* 50:82–92
- Hu Q, Zhou T (2018) ElCiRNA-mediated gene expression: tunability and bimodality. *FEBS Lett* 592:3460–3471
- Huh D, Paulsson J (2011) Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat Genet* 43:95–100
- Itakura AK, Chakravarty AK, Jakobson CM, Jarosz DF (2020) Widespread prion-based control of growth and differentiation strategies in *Saccharomyces cerevisiae*. *Mol Cell* 77(266–278):e266
- Keren L, van Dijk D, Weingarten-Gabbay S, Davidi D, Jona G, Weinberger A, Milo R, Segal E (2015) Noise in gene expression is coupled to growth rate. *Genome Res* 25:1893–1902
- Kim S, Jacobs-Wagner C (2018) Effects of mRNA degradation and site-specific transcriptional pausing on protein expression noise. *Biophys J* 114:1718–1729
- Kinkhabwala A, Khmelinskii A, Knop M (2014) Analytical model for macromolecular partitioning during yeast cell division. *BMC Biophys* 7:10
- Levy SF, Ziv N, Siegal ML (2012) Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS Biol* 10:e1001325
- Liu J, Francois JM, Capp JP (2019) Gene expression noise produces cell-to-cell heterogeneity in eukaryotic homologous recombination rate. *Front Genet* 10:475
- Losick R, Desplan C (2008) Stochasticity and cell fate. *Science* 320:65–68

- Maheshri N, O'Shea EK (2007) Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu Rev Bioph Biom* 36:413–434
- Makova KD, Hardison RC (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* 16:213–223
- McDonald MJ (2019) Microbial experimental evolution—a proving ground for evolutionary theory and a tool for discovery. *EMBO Rep* 20:e46992
- Mikl M, Hamburg A, Pilpel Y, Segal E (2019) Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat Commun* 10:4572
- Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441:840–846
- O'Dea RE, Noble DWA, Johnson SL, Hesselson D, Nakagawa S (2016) The role of non-genetic inheritance in evolutionary rescue: epigenetic buffering, heritable bet hedging and epigenetic traps. *Environ Epigenet* 2:dvv014
- Paulsson J, Berg OG, Ehrenberg M (2000) Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc Natl Acad Sci U S A* 97:7148–7153
- Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. *Science* 307:1965–1969
- Phillips NE, Manning CS, Pettini T, Biga V, Marinopoulou E, Stanley P, Boyd J, Bagnall J, Paszek P, Spiller DG et al (2016) Stochasticity in the miR-9/Hes1 oscillatory network can account for clonal heterogeneity in the timing of differentiation. *eLife* 5:e16118
- Raj A, Rifkin SA, Andersen E, van Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. *Nature* 463:913–918
- Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304:1811–1814
- Raser JM, O'Shea EK (2005) Noise in gene expression: origins, consequences, and control. *Science* 309:2010–2013
- Roberts CWM, Leroux MM, Fleming MD, Orkin SH (2002) Highly penetrant, rapid tumorigenesis through conditional inversion of the tumor suppressor gene *Snf5*. *Cancer Cell* 2:415–425
- Ruess J, Pleska M, Guet CC, Tkacik G (2019) Molecular noise of innate immunity shapes bacteriophage ecologies. *PLoS Comput Biol* 15
- Sanchez A, Choubey S, Kondev J (2013) Regulation of noise in gene expression. *Annu Rev Biophys* 42:469–491
- Schmiedel JM, Klemm SL, Zheng Y, Sahay A, Bluthgen N, Marks DS, van Oudenaarden A (2015) Gene expression. MicroRNA control of protein expression noise. *Science* 348:128–132
- Siegal ML, Leu JY (2014) On the nature and evolutionary impact of phenotypic robustness mechanisms. *Annu Rev Ecol Evol Syst* 45:495–517
- Simons AM (2011) Modes of response to environmental change and the elusive empirical evidence for bet hedging. *Proc Biol Sci* 278:1601–1609
- Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34:D535–D539
- Stewart-Ornstein J, Weissman JS, El-Samad H (2012) Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. *Mol Cell* 45:483–493
- Taubner J, Wieczorek D, Yasin L, Brozou T, Borkhardt A, Kuhlen M (2018) Penetrance and expressivity in inherited cancer predisposing syndromes. *Trends Cancer* 4:718–728
- Uphoff S, Lord ND, Okumus B, Potvin-Trottier L, Sherratt DJ, Paulsson J (2016) Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science* 351:1094–1097
- Weinberger L, Voichek Y, Tirosch I, Hornung G, Amit I, Barkai N (2012) Expression noise and acetylation profiles distinguish HDAC functions. *Mol Cell* 47:193–202
- Wernet MF, Perry MW, Desplan C (2015) The evolutionary diversity of insect retinal mosaics: common design principles and emerging molecular logic. *Trends Genet* 31:316–328

- Wiesenfeld K, Moss F (1995) Stochastic resonance and the benefits of noise: from ice ages to crayfish and SQUIDS. *Nature* 373:33–36
- Wu S, Li K, Li Y, Zhao T, Li T, Yang YF, Qian W (2017) Independent regulation of gene expression level and noise by histone modifications. *PLoS Comput Biol* 13:e1005585
- Yona AH, Frumkin I, Pilpel Y (2015) A relay race on the evolutionary adaptation spectrum. *Cell* 163:549–559
- You ST, Zhou YT, Kao CF, Leu JY (2019) Experimental evolution reveals a general role for the methyltransferase Hmt1 in noise buffering. *PLoS Biol* 17:e3000433