Garrido-Ramos MA (ed): Repetitive DNA.
Genome Dyn. Basel, Karger, 2012, vol 7, pp 170–196

The Birth-and-Death Evolution of Multigene Families Revisited

J.M. Eirín-López[a] · L. Rebordinos[b] · A.P. Rooney[d] · J. Rozas[c]

[a]CHROMEVOL-XENOMAR Group, Departamento de Biología Celular y Molecular, Universidade da Coruña, A Coruña, [b]Área de Genética, Facultad de Ciencias del Mar y Ambientales, Universidad de Cádiz, Cádiz, [c]Departament de Genètica y Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain; [d]Crop Bioprotection Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, US Department of Agriculture, Peoria, Ill., USA

Abstract

For quite some time, scientists have wondered how multigene families come into existence. Over the last several decades, a number of genomic and evolutionary mechanisms have been discovered that shape the evolution, structure and organization of multigene families. While gene duplication represents the core process, other phenomena such as pseudogene formation, gene loss, recombination and natural selection have been found to act in varying degrees to shape the evolution of gene families. How these forces influence the fate of gene duplicates has ultimately led molecular evolutionary biologists to ask the question: How and why do some duplicates gain new functions, whereas others deteriorate into pseudogenes or even get deleted from the genome? What ultimately lies at the heart of this question is the desire to understand how multigene families originate and diversify. The birth-and-death model of multigene family evolution provides a framework to answer this question. However, the growing availability of molecular data has revealed a much more complex scenario in which the birth-and-death process interacts with different mechanisms, leading to evolutionary novelty that can be exploited by a species as means for adaptation to various selective challenges. Here we provide an up-to-date review into the role of the birth-and-death model and the relevance of its interaction with forces such as genomic drift, selection and concerted evolution in generating and driving the evolution of different archetypal multigene families. We discuss the scientific evidence supporting the notion of birth-and-death as the major mechanism guiding the long-term evolution of multigene families.

Forty-one years ago Susumu Ohno [1] stated that gene and genome duplications are the major evolutionary mechanisms for generating functional innovation. Since then, we have learned much regarding the evolutionary processes that influence nucleotide and amino acid substitution, both at the intraspecific and interspecific levels [2]. However, our current understanding of gene duplication dynamics is considerably

less [3]. Despite the fact that a number of models and hypothesis have been developed to describe the evolutionary dynamics of gene duplications within and between species, the lack of readily available, high quality data limited our ability to test the applicability of most models to real data in past studies of the 'pre-genomic' era. The 2 main sources of problems were (1) the lack of complete genome information for many, if not most, gene families, and (2) the lack of accurate methods for inferring orthologous-paralogous gene relationships [4].

Gene families can be classified according to a number of criteria [3, 5, 6]. Such criteria may include, for example, (1) function, (2) how members are distributed across the genome, and (3) the primary mechanism responsible for generating the families in question. For instance, gene families have been categorized separating those organized into gene clusters from those with members at dispersed locations across the chromosomes. Yet, a classification based on the underlying mechanism for the origin of the family members is, in many cases, much more informative: not only does it explain the chromosomal distribution of family members, but it also provides insights into their evolutionary fate. Gene families essentially arise by 2 basic gene duplication mechanisms: unequal crossing-over and retroposition [7]. The first mechanism usually creates tandem repeats physically linked on the chromosomes and, therefore, in a non-random fashion. The family members in this case may have introns (if the original gene had introns) and non-coding regulatory sequences. In contrast, retroposition results in the insertion of an intronless cDNA with losses of upstream non-coding regions and with poly(A) tracts, more or less at random, at locations dispersed across the genome. The knowledge of the mechanism of origin is critical to understanding the forces that drive the generation of gene clusters; for example, a particular cluster of genes might have arisen simply due to random chance of having been located in a region of the genome more prone to unequal crossing-over than in other regions.

The recent availability of complete genomes from closely related species has provided valuable opportunities to conduct extensive studies of gene family evolution [8]. The analyses of these new data, however, also present a number of difficulties that remain to be solved such as, for example, the inability of current assembling algorithms to handle highly repetitive DNA sequences. Another problem concerns the accurate inference of orthologous-paralogous relationships. Currently, gene gain and loss events can be estimated either from the number of gene family members in the extant species of a phylogeny [9, 10], or via gene tree/species tree reconciliation [11]. The latter methods, however, have important limitations [8], such as their dependence on the correct gene tree and the true species tree, as well as the incomplete lineage sorting problem. Although there have been some improvements to minimize the gene tree uncertainty by taking into account clade support values, branch lengths [12] or synteny information [13], gene tree/species tree reconciliation is not well suited in order to conduct statistical hypothesis testing, and as such, it has limitations in its application.

Models of Multigene Family Evolution

The study of the mechanisms governing the evolution of multigene families has constituted a controversial issue ever since sets of functionally related genes were first discovered. The aforementioned limitations and others, such as the lack of detailed knowledge pertaining to the structure, organization and diversity of family members, their functional meaning as well as the lack of accurate methodologies for determining phylogenetic homology among sequences have fueled this controversy. The first efforts focused on deciphering the evolutionary dynamics of gene families date back to the early 1960s, with studies using hemoglobin and myoglobin as model systems [14]. The finding that the genes encoding these proteins are phylogenetically related and that they acquired new gene functions through their gradual divergence led to the proposal of the first general model of evolution of these multigene families, referred to as 'divergent evolution'.

The validity of the divergent evolution model was quickly challenged by the growing amount of data collected from studies on additional families, especially those displaying tandemly arrayed organizations (i.e. ribosomal DNA (rDNA) and histones). Within this context, the development of DNA sequencing techniques during the 1970s helped researchers to analyze the patterns of variation in coding and noncoding regions, unveiling that nucleotide sequences of different multigene family members are more closely related within species than between species. Such observations (which deviate from the predictions made by the divergent evolution model) were explained by an alternative model of multigene family evolution termed 'concerted evolution'. According to this model, after the split of an ancestral species into 2 descendent ones, the members of a repeated gene family would evolve together as a block, displaying a high degree of homogeneity within a given descendant species as they gradually diverged with respect to repeats from closely related species. Under this model, sequence homogenization results from random unequal crossing-over and gene conversion among gene family members, although some gene variants are expected to occur due to mutation.

The apparent efficiency of the concerted evolution model in explaining the observed patterns of molecular variation quickly overshadowed any alternative explanation throughout the 1970s and 1980s, consolidating the notion that most multigene families evolve following this model. Indeed, it was not until the early 1990s that concerted evolution began to be seriously questioned, especially as a result of the growing availability of molecular data coming from the dawn of the genomic era. Surprisingly, these data revealed that far from being conserved and homogeneous, most multigene families encompassed far too much intraspecific diversity (genetic and functional) to be consistent with a homogenizing mechanism. These conclusions, together with other atypical features observed across multigene family members (most notably the presence of between-species clustering patterns in phylogenies and the presence of pseudogenes), motivated the proposal of a new model termed 'birth-and-death evolution' [15]. In contrast to the concerted evolution model, the birth-and-death model

promotes genetic diversification and provides an explanation for the generation of new gene families.

The Birth-and-Death Model of Evolution

Over the last 2 decades, Nei and colleagues conducted a number of key studies that provide the foundation for the theory that underlies the birth-and-death model. Since then, a number of multigene families have been identified that undergo birth-and-death evolution (reviewed in [6]). The basic foundational elements of the model are the differential levels of gene duplication and subsequent loss or maintenance of gene copies within a multigene family. Accordingly, when duplication gives rise to new copies of a gene, and these copies do not evolve in concert as discussed in the previous section, some of the copies may persist in the genome for long periods of time. Eventually, the copies diverge in sequence such that they no longer are identical nor do they possess extensive regions of similarity. On the other hand, some copies of the original 'parent' gene may degenerate into pseudogenes or they may get deleted from the genome through, for example, unequal crossing-over. Consequently, the most common way to determine if birth-and-death evolution characterizes a multigene family is to look for the 2 hallmark features of the model: (1) an interspecific gene clustering pattern and (2) the presence of pseudogenes.

There are cases in which an interspecific phylogenetic clustering pattern and/or pseudogene formation are not detectable. While the latter is dependent mostly on intrinsic genome dynamics and random chance, both are dependent upon proper analytical techniques. Still there are instances in which even thorough, proper analyses can lead to the erroneous conclusion that birth-and-death evolution does not occur, simply because an intraspecific gene clustering pattern was observed in the reconstructed multigene family phylogeny. Such false conclusions can arise from (1) recent gene duplication within a species; (2) strong purifying selection; and (3) rapid gene turnover. With respect to recent gene duplication, enough nucleotide substitutions will accumulate over time such that divergence between gene duplicates eventually becomes detectable (albeit over hundreds of thousands, if not millions, of years). Thus, a pattern of between-species gene clustering will characterize the phylogeny of the multigene family, provided that enough time has elapsed for the divergence of gene duplicates and/or their orthologs present in different genomes [6]. In cases involving strong purifying selection, one must consider the differences in the way in which substitutions accumulate and are distributed between protein-coding and non-protein-coding genes. For protein-coding genes, an analysis of divergence levels at synonymous versus non-synonymous sites will reveal if purifying selection, and not concerted evolution, is the cause for sequence constraint; indeed, under purifying selection synonymous sites will have some divergence levels even when non-synonymous sites show no variation [16]. In the case of genes that do not encode protein, such as ribosomal RNA (rRNA) genes, the analysis is more difficult and often requires study of nucleotide substitution levels in the regions immediately flanking the genes as well as in introns or intergenic spacer

regions, if present, followed by comparison to sequence divergence within the coding region of the gene [17]. Differential levels of nucleotide sequence conservation between the coding and non-coding regions may reveal if purifying selection is the determinant for sequence conservation. Finally, when rapid gene turnover occurs within a multigene family, deletion and duplication are so frequent that orthologous gene pairs are quickly lost between species, so a within-species clustering pattern predominates [17–20]. But some amount of nucleotide substitution should still be observable and there may be at least some between-species clustering events, both of which are indicators that birth-and-death evolution has occurred. The aforementioned examples are only a few, and there may be more that can produce potentially misleading results in analyses designed to detect birth-and-death evolution.

It should be noted that, while considerable effort has been given to the study of gene duplication, little attention has been paid to the effects of gene deletion on multigene family evolution. Thus, much like the failure to recognize recent gene duplication, strong purifying selection and rapid gene turnover as causes of within-species gene clustering patterns, the failure to recognize the importance of gene loss may result in phylogenetic patterns that could be misinterpreted as, for instance, lateral transfer events [21]. In this case, however, phylogenetic analyses may not help and the problem is rendered intractable (see [21] for a review of this topic).

There are a number of genomic and evolutionary mechanisms that can shape the structure, organization and evolution of multigene families (see [6]). For the last decades, concerted evolution has prevailed as the 'default' long-term evolutionary model for the evolution of most (if not all) multigene families. We nowadays know that multigene families encompass too much genetic diversity to be generated and maintained by means of such a homogenizing mechanism. Indeed, comprehensive studies conducted during the last 10 years, addressing the evolution of multigene families, usually support the birth-and-death process as the underlying mechanism. However, in spite of the evidence gathered in favor of this latter model, birth-and-death has only shyly replaced concerted evolution as the 'default' model of long-term evolution of multigene families. In the present chapter we provide an up-to-date review into the role of the birth-and-death model and its interaction with forces such as genomic drift, natural selection and concerted evolution in generating and driving the evolution of different archetypal multigene families. Here we show empirical evidence supporting the concept of birth-and-death as the major mechanism underlying the long-term evolution of multigene families.

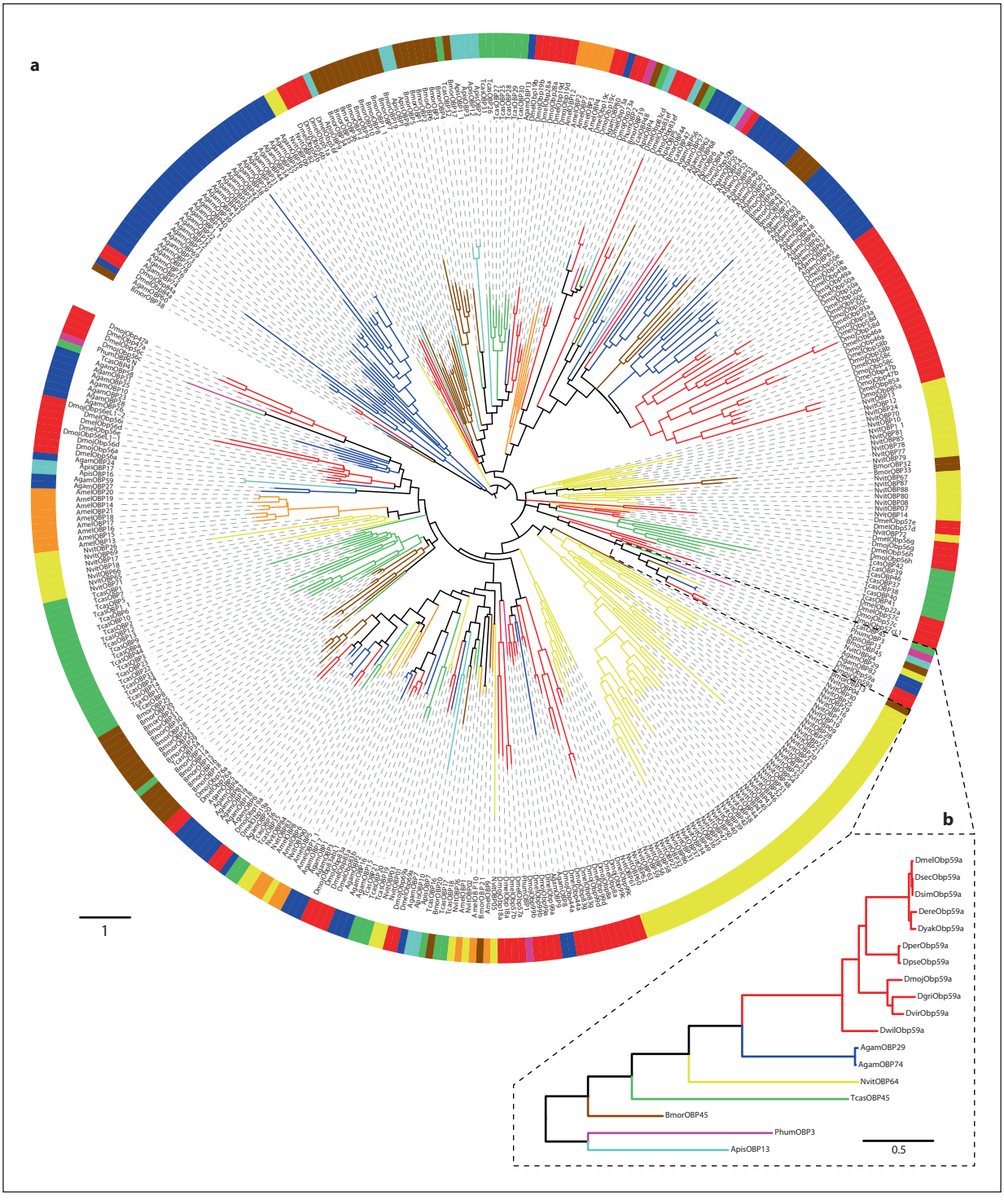
Rates of Birth-and-Death Evolution: Lessons from Gene Families of the Chemosensory System

Despite the availability of complete genome information for a number of eukaryotic species, we are far from understanding the forces that have driven the lineage-specific

expansions (or contractions) of many multigene families, as well as more general features that characterize their evolution. The most important limitations are the following: (1) quite often, so-called complete genomes are not fully completed and are very fragmented. This is a very important problem since repetitive DNA regions are usually the worst assembled and, therefore, often incompletely represented in a genome annotation; this limitation is more critical for tandemly distributed repetitive regions than for those showing a more dispersed distribution across the genome. Hence, we lack detailed information (number of copies, physical location on the chromosomes) for many gene families, but especially those that exist in large clusters of tandemly repeated genes. (2) Many species for which completed genomes are available are separated by vast evolutionary times. Indeed, there are few cases in which we have genome information from relatively closely related species (e.g. within a genus or within a family); the genome sequence of 12 *Drosophila* species is one of such few examples [22]. Since many gene families have relatively high gene turnover rates (birth-and-death rates), information from highly divergent genomes can confound fine and exhaustive lineage-specific analyses (e.g. the accurate determination of the numbers of gene gains and losses might be highly inaccurate depending upon the rate of gene turnover). (3) Current methods for inferring orthologous-paralogous relationships may have low accuracy [4] (e.g. gene tree and species tree problem), when gene conversion is frequent or when large numbers of gene gains and losses have occurred. Regardless, limitation 2 likely will no longer be a problem in the near future, but limitations 1 and 3 may take longer to be resolved.

A comparative genome analysis using the complete set of genes in a phylogenetic framework provides the most conclusive evidence on the gene family's origin and evolutionary fate. In particular, analyses including genomes from closely and distantly related species have been shown to constitute a very successful approach. The genome analyses of the major gene families involved in the chemosensory system of the insects represent a good example to illustrate the state-of-the-art of gene family evolutionary analysis using complete genome DNA sequence data [13, 20, 23]. The most important proteins implicated in the early chemoreception steps in insects are encoded by gene families of moderate size. This process, which occurs inside the aqueous fluid of the chemosensory hair-like structures named sensilla, comprises the first contact of the external chemical signals (the odorant in the olfactory system) with membrane chemoreceptor proteins (the olfactory receptors in the olfactory system).

These multigene families can be classified into 2 main functional groups, the odorant-binding (OBPs) and chemosensory (CSPs) proteins (involved in the transport of the chemical signals through the sensillar lymph), and the chemosensory receptors that recognize the external cues and translate this information into an electrical signal (a dendritic spike) to the central nervous system, which elicit the appropriate behavior. In insects, there are 3 chemosensory receptor gene families: the olfactory (ORs) and gustatory (GRs) receptors, which in turn encompass the chemoreceptor superfamily, and the ionotropic receptors (IRs). Comparative


a
1
b
DmelObp59a
DsecObp59a
DsimObp59a
DereObp59a
DyakObp59a
DperObp59a
DpseObp59a
DmojObp59a
DgriObp59a
DvirObp59a
DwilObp59a
AgamOBP29
AgamOBP74
NvitOBP64
TcasOBP45
BmorOBP45
PhumOBP3
ApisOBP13
0.5

genome analyses revealed that the size of these multigene families differs markedly across species [20, 23] (fig. 1). While the number of genes of the OBP family ranges from 21 (in *Apis mellifera*) to 83 (in *Anopheles gambiae*), the CSP numbers range from 3 (in *Drosophila ananassae*) to 22 (in *Bombyx mori*); and whereas the ORs vary from 48 (in *Acyrthosiphon pisum* and in *B. mori*) to 265 (in *Tribolium castaneum*), the GRs range from 10 *(A. mellifera)* to 220 *(T. castaneum)*, and the IRs from 10 *(A. mellifera)* to 95 *(Aedes aegypti)*. Furthermore, these figures do not include information for the body louse *(Pediculus humanus)*, which contains a considerably lower number of genes (5 OBPs, 7 CSPs, 10 ORs, 8 GRs and 12 IRs), the cause for which likely stems from its parasitic lifestyle. This disparate number of genes in different insect species, nevertheless, provides a good opportunity to gain insight into the evolutionary mechanisms shaping gene family sizes and, particularly, into the role of natural selection and adaptation. Furthermore, the fact that these gene families include a moderate number of members allows for a comprehensive analysis that combine both automatic and manual 'gene calling' efforts, and also increases the accuracy of the resulting annotation.

It has been shown that the major gene families of the chemosensory system are usually arranged in chromosome clusters [23]. For instance, nearly 70% of the *Drosophila melanogaster* OBP genes (52 genes) are arranged in 10 clusters of 2–6 genes each. Nevertheless, despite the fact that this kind of arrangement also exists in other insect species and in other gene families, the actual fraction of the genes arranged in clusters is highly variable. Interestingly, physically neighboring members of these families are also phylogenetically related; for instance, evolutionarily new OBP duplicates are usually identified in extant chromosomal clusters, whereas phylogenetically close OBP genes are also located in the same cluster. Such data clearly supports unequal crossing-over as the main mechanism that generates tandem gene duplications of the chemosensory gene families.

Phylogenetic Analyses and the Birth-and-Death Process

Phylogenetic analyses including orthologous and paralogous copies show that the actual number of members is relatively conserved across the *Drosophila* genus, with few examples of species-specific expansions. However, a fine-scale investigation

Fig. 1. Phylogenetic analysis of the insect OBP genes. **a** Amino acid sequences of *A. gambiae* (Agam), *A. mellifera* (Amel), *A. pisum* (Apis), *B. mori* (Bmor), *D. melanogaster* (Dmel), *D. mojavensis* (Dmoj), *Nasonia vitripennis* (Nvit), *P. humanus* (Phum), and *T. castaneum* (Tcas). **b** Phylogenetic relationships of the OBP59a orthologous group in species of panel **a** and the following *Drosophila* species: *D. erecta* (Dere), *D. grimshawi* (Dgri), *D. persimilis* (Dper), *D. pseudoobscura* (Dpse), *D. sechellia* (Dsec), *D. simulans* (Dsim), *D. virilis* (Dvir), *D. willistoni* (Dwil), and *D. yakuba* (Dyak) . The *OBP59a* gene is absent in *A. mellifera*. The phylogenetic branches (and the outer ring) of the different species are depicted in colors: red, *Drosophila* species; blue, *A. gambiae*; brown, *B. mori*; green, *T. castaneum*; orange, *A. mellifera*; yellow, *N. vitripennis*; cyan, *A. pisum;* and pink, *P. humanus*. The scale bar represents 1 (**a**) or 0.5 (**b**) amino acid substitutions per site.

uncovers a large number of gene gains, gene losses and pseudogenization events, although these events have different frequency among gene families. Noticeably, gene losses and pseudogenization events are unequally distributed across the *Drosophila* phylogeny; indeed, the later events are mainly inferred in the terminal branches, suggesting that pseudogenes have a very short half-life. Across this genus, furthermore, it is reasonably easy to observe orthologous groups including all *Drosophila* species and, for a particular orthologous group there usually exists a good reconciliation between gene and species trees (fig. 1b). This data strongly suggests that these genes have diverged independently since their origin. These figures, however, are different from those found when distantly related species are compared (e.g. between insect orders) (fig. 1a). Indeed, there is a dramatic variation in gene family size as well as few examples of genes with orthologous copies across insects and many lineage-specific gene expansions (fig. 1a). Both features, however, are caused by the same basic evolutionary mechanism, the birth-and-death model (see below).

Current analyses of the chemosensory gene families (mostly from the OBP gene family data) within a phylogenetic framework largely support the birth-and-death model of evolution [6], specifically: (1) several gene gain and loss events have occurred in the evolution of the gene family; (2) a number of nonfunctional members (pseudogenes) can be identified across the phylogeny (mostly in terminal phylogenetic branches); (3) the phylogenetic trees inferred from orthologous genes fit well with the accepted species phylogeny; (4) there is no evidence for a major impact of gene conversion in the evolution of paralogous genes (although current methods for detecting gene conversion may be insufficient); (5) the number of orthologous groups including representatives of all surveyed species gradually decreases with increasing divergence time; (6) there is an uneven phylogenetic subfamily distribution across species; and (7) several gene expansions and contractions are identified across large (e.g. within-class or within-order) but not across short (e.g. across a genus) evolutionary times.

Birth-and-Death Rates and the Impact of Natural Selection

Methods and software have been developed to estimate birth-and-death rates (e.g. [13, 24]). The CAFE software [24] implements a stochastic birth-and-death model which allows an estimation of birth-and-death rates using a maximum likelihood approach (λ is the birth-and-death rate per gene and per million years) under the assumption of equal birth-and-death rates. Although this assumption may not always hold (e.g. in the presence of family expansions), it is a useful method for comparing birth-and-death rates across gene families or across species. For example, the birth-and-death rate for the complete set of gene families of *Drosophila* has been estimated as $\lambda = 0.0012$ [25], which indicates that there have been ~17 new gene gains or ~17 losses every million years during the evolution of any one *Drosophila* species' genome. In addition, the birth-and-death rates for the chemosensory gene families are noticeably larger than the estimates for the complete *Drosophila* genomes (OBPs, $\lambda = 0.005$; ORs, $\lambda = 0.006$; GRs, $\lambda = 0.011$; IRs, $\lambda = 0.0023$) [13, 23]; for instance, the

value of λ = 0.005 inferred for the OBP gene family (assuming ~50 members) suggests that there has been an OBP gene gain (or a loss) every 4 million years. Such features, therefore, indicate that these gene families have a highly dynamic mode of evolution through which new members are continuously counterbalancing gene losses or non-functionalizations and pseudogenizations.

These high gene turnover rates exhibited by the chemosensory gene families additionally are shaped by natural selection. Indeed, natural selection can modify the rate of fixation in the population of newly duplicated copies, and it also can contribute to the functional diversification associated with sequence divergence. The levels of functional constraint and functional divergence can be analyzed through the comparative analysis of the ratio of non-synonymous (d_N) to synonymous (d_S) divergence ($\omega = d_N/d_S$), in which the ω value serves as a proxy for gauging levels of functional constraint. This method allows for the quantification of the impact of purifying (negative) and adaptive (positive) selection as well as for the testing of contrasting alternative evolutionary hypotheses. In the absence of selection the expected value of ω is 1, whereas statistically significant values lower (or higher) than 1 might be indicative of purifying (or positive) selection. The ω estimates for the OBPs, ORs and GRs of *Drosophila* clearly point to purifying selection as the main evolutionary force (OBPs, ω = 0.15; ORs, ω = 0.14; GRs, ω = 0.22). These ω values, furthermore, differ significantly among genes within a particular gene family. For instance, the ω values among the OBP orthologous groups range from 0.003 to 0.11. Among the ORs, the *Obp83b* gene has the smallest ω ratio, which is consistent with its critical function and its strong conservation across the insects. There are also strong differences among GR members; for instance, the sweet taste and the carbon dioxide receptors display low ratios. The functional constraint levels can also vary across positions of the coding region. Indeed, the specific molecular fingerprint of positive selection could even be detected in amino acids located in the putative odorant binding pocket of some OBPs. Since these changes likely affect the sensitivity or specificity in detecting odorants, these regions may be more likely to evolve by positive selection.

Birth-and-Death Evolution and Genomic Drift: Evolving Evolutionary Novelty in the Fatty Acid Reductase Multigene Family

During the evolutionary history of a multigene family that evolves under a birth-and-death model, the random occurrence of gene duplication and loss can lead to a change in the number of gene copies (i.e. dosage repetition) or paralogous family members (i.e. variant repetition) present within a genome. Thus, if one tallies the number of gene copies or family members present in a species' genome and compares it to a different species' genome, the numbers may be different. Nei [26] termed this 'genomic drift' and likened it to the random change of allele frequencies at a single gene produced by genetic drift.

For the most part, one expects the number of genes that are present in a genome to be determined solely through random chance. Dosage repetition, however, is one instance in which selection may play a role in determining gene copy number. For example, it is generally accepted that a large number of rRNA gene copies facilitates mRNA transcription, and therefore there exists a lower limit on the number of copies that a genome will tolerate. Consequently, the bobbed mutant phenotype of *D. melanogaster* appears when there is a loss of 50% or more of wild-type rRNA genes; in cases in which less than 15% of the wild-type rRNA genes remain, the mutation is lethal [27]. Likewise, adaptation to a novel environment or set of ecological circumstances can also drive changes in gene copy number [28]. For example, the evolution of tetrapods from a fish ancestor was accompanied by a concomitant increase in the number of paralogous olfactory genes present in the ancestral tetrapod genome, presumably in response to the increased number of odorants found on land versus in the aquatic environment [28]. Accordingly, once these new gene duplicates began to diverge from their parental gene, novel functions were acquired and, presumably, the number of odorants that could be detected subsequently increased.

The extent to which genomic drift influences a multigene family can be studied through the inference of the number of gene duplication and loss events that have occurred during the evolutionary history of the family [28, 29]. This is accomplished through the 'reconciliation' of the gene tree (i.e. the multigene family phylogeny) with the species tree [30–33]. In short, this procedure involves inferring the lowest number of duplication and loss events required to produce the observed gene tree given the assumed species tree. The procedure is too laborious to carry out by hand even when there are relatively small numbers of paralogous gene copies; thus, the use of computer software to conduct these analyses is highly recommended (e.g. NOTUNG [30]). To demonstrate how such an analysis is conducted, below we present a case study of the fatty acyl-coenzyme A reductase, or fatty acid reductase (FAR), multigene family using sequences extracted from the complete genomes of representative species of eukaryotes (fig. 2).

Genomic Drift Between Multigene Families

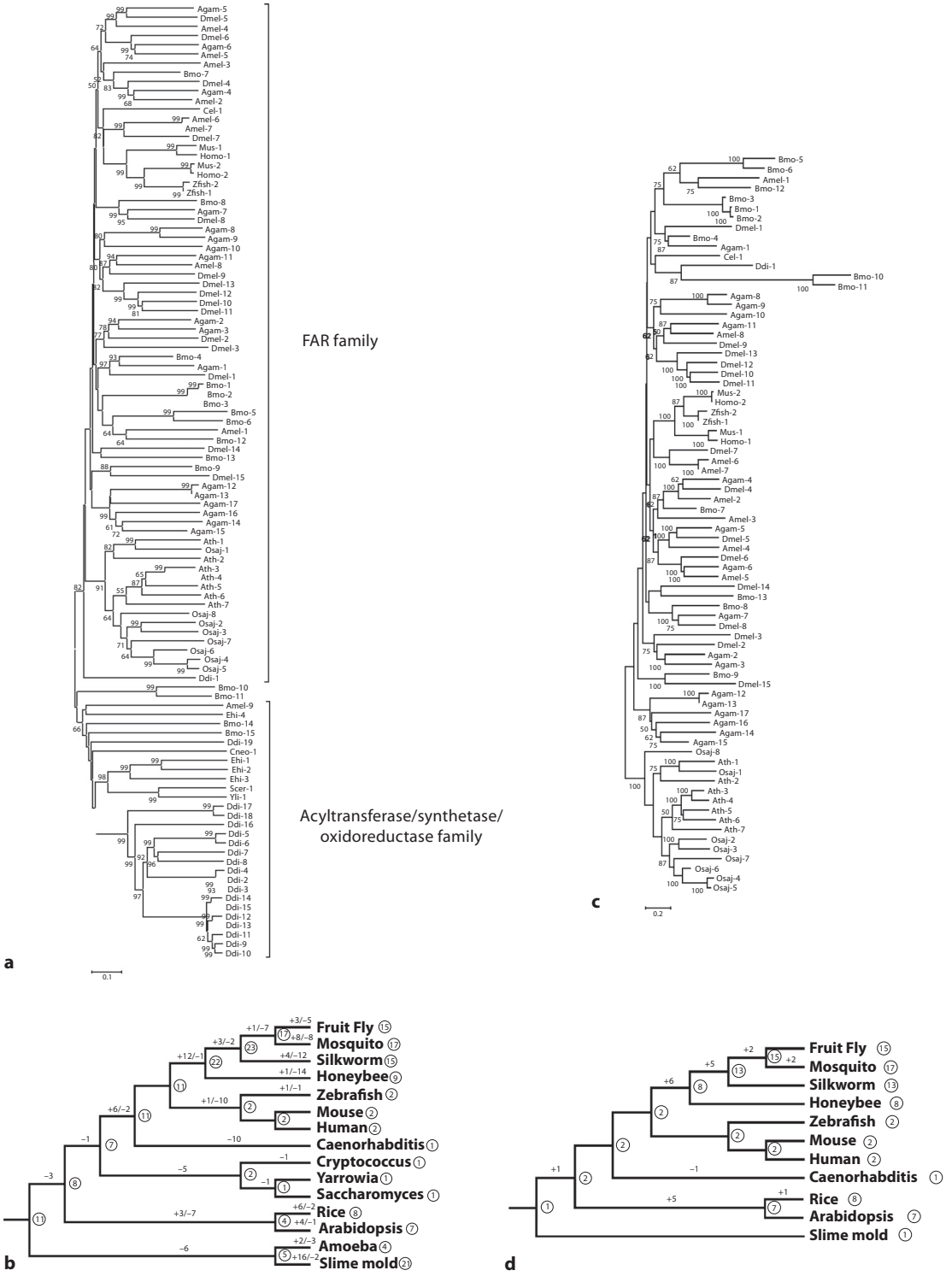
FAR enzymes catalyze the reduction of fatty acids to fatty alcohols in a reaction that is dependent upon NADPH as a cofactor. The number of FAR genes per genome can vary greatly between organisms. In vertebrates, there are 2 reductase genes present in the genome, whereas there are more than a dozen present in the silkworm. The evolutionary origins of this gene family are not well understood, but we found that acyl-CoA synthetase, acyltransferase and oxidoreductase gene families are close relatives of this family on the basis of protein sequence similarity (data not shown) and thus form a superfamily. If we examine the phylogenetic relationships of representatives of this superfamily (fig. 2a), we see evidence of birth-and-death evolution as shown through a pattern of between-species gene clustering. There are a couple of instances in which large single-species gene clusters were found (e.g. slime mold genes). This,

however, is not unexpected since there is a lack of a closely related species to include in the comparison in this case. There is no evidence for concerted evolution of these genes, as the branch lengths found within these clusters are all relatively long, indicating that at least a moderate amount of divergence has occurred.

We can examine the question of gene turnover dynamics in more detail through an analysis of gene gain and loss. The analysis shown in figure 2b reveals that a varying amount of activity has taken place over the evolution of this superfamily. At the root of the phylogeny, which represents the last common ancestor shared between the 'lower' (i.e. slime mold and amoeba) and 'higher' eukaryote representatives studied, the ancestral genome was inferred to have possessed 11 genes constituting this superfamily. A considerable amount of gene turnover can be inferred to have occurred as shown through the different numbers of genes present in the various ancestral genomes (internal nodes) in the phylogeny (fig. 2b). Of particular interest is the observation that insects gained substantially higher amounts of genes than the other lineages, whereas vertebrates and nematodes (as represented by *Caenorhabditis elegans*) lost substantially more. As the FAR gene family dominates this superfamily in terms of total numbers of genes (fig. 2a), we can assume that most of this activity involves that family. To test this hypothesis and possibly determine the cause for the pattern, we conducted a separate analysis of the FAR multigene family.

Genomic Drift Within a Multigene Family

The results of our analysis of the FAR multigene family are presented in figures 2c and 2d. Expectedly, the FAR gene family undergoes birth-and-death evolution (fig. 2c) in accordance with the pattern inferred in figure 2a for the superfamily as a whole. However, the pattern of gene gain and loss is substantially different (fig. 2d). Virtually no gene loss was found to have occurred since the divergence of the slime mold from 'higher' eukaryotes (fig. 2d). In fact, the only lineage in which gene loss was found to happen was the nematode lineage (as represented by *C. elegans*), which involved the loss of only a single gene. In contrast, the pattern of gene gain is more dynamic. Two bursts are notable: (1) plants apparently gained 5 genes since they diverged from their last common ancestor shared with animals (fig. 2d), and (2) insects gained a substantial number of genes since they diverged from other animals: 6 genes were gained after they diverged from their last common ancestor shared with vertebrates, and another 5 genes were gained after the divergence of the honeybee from the silkworm and the fruit fly and mosquito. However, the gain of 6 genes along the lineage leading to insects from their last common ancestor with vertebrates must be interpreted with some caution, because there are a substantial number of other insect orders that are not represented in this phylogeny as well as other invertebrate and vertebrate lineages. Consequently, this number could be the result of 'summing' across other internal branches not found within this phylogeny due to missing taxa. This caveat may also hold for the number of gains along the branch leading to plants. In contrast, the gain of 5 genes in the common ancestor of


FAR family
Acyltransferase/synthetase/
oxidoreductase family
a
b
c
d
Fruit Fly 15
Mosquito 17
Silkworm 15
Honeybee 9
Zebrafish 2
Mouse 2
Human 2
Caenorhabditis 1
Cryptococcus 1
Yarrowia 1
Saccharomyces 1
Rice 8
Arabidopsis 7
Amoeba 4
Slime mold 21
Fruit Fly 15
Mosquito 17
Silkworm 13
Honeybee 8
Zebrafish 2
Mouse 2
Human 2
Caenorhabditis 1
Rice 8
Arabidopsis 7
Slime mold 1

the silkworm, fruit fly and mosquito subsequent to their divergence from the honeybee is likely more reliable, since there are fewer missing taxa relative to the taxonomic rank (order) represented by the species in the study and, therefore, unlikely to alter the number much.

Regardless, simple calculation of the number of FAR genes present in the genomes of the species studied clearly indicates that plants and insects have undergone large expansions relative to the other taxa examined, which is consistent with the genomic drift hypothesis of Nei [26] for multigene families undergoing birth-and-death evolution. The possibility that these expansions facilitated the adaptive evolution of a variety of specialized functions that involve precursors upon which FAR genes act is especially interesting. For example, FAR genes have been shown to function in pheromone biosynthesis in moth species directly through the production an alcohol that confers species-specificity or indirectly through the biosynthesis of precursor compounds [34]. If we can assume that the silkworm is truly representative of moth species, the large number of FAR genes present in moth genomes (13 in the silkworm; fig. 2d) and the variety in substrate specificity that these genes have been shown to display [34, 35] suggest a number of different specialized functions have evolved. Similarly, plant FAR genes also have been shown to have evolved a number of specialized functions, such as the biosynthesis of wax esters used for storage in developing seeds [36], the biosynthesis of the lipid component used in the outer pollen wall, and the biosynthesis of cuticular wax lipids [37]. In contrast, the other species studied have very few FAR genes or even no genes. For example, *C. elegans* and the slime mold were found to have only 1 gene, and vertebrates only have 2, whereas the 3 fungi *(Cryptococcus neoformans, Yarrowia lipolytica*, and *Saccharomyces cerevisiae)* and the amoeba *(Entamoeba histolytica)* did not have any FAR homologues. It is possible that these species rely less on FAR genes to synthesize the fatty alcohol-containing compounds that these species require and other genes have evolved to take over these functions, or perhaps these species

Fig. 2. Birth-and-death evolution of FAR genes. **a**, **b** Phylogenetic analysis of the FAR/acyltransferase/oxidoreductase superfamily (**a**) and the associated gene tree reconciliation analysis for the superfamily (**b**). **c**, **d** Phylogenetic analysis of the FAR multigene family (**c**) and the associated gene tree reconciliation analysis for this family (**d**). **a**, **b** The computer program MEGA 4 [65] was used to reconstruct trees from Poisson amino acid distances using the neighbor-joining method. Numbers along branches represent bootstrap percentage values generated from 1,000 pseudoreplicates; only numbers greater than 50% are shown. **c**, **d** The computer program NOTUNG 2.6 [30] was used to conduct gene tree reconciliation analyses. The phylogenies shown are species trees based on [66]. Numbers along branches denote gene gains (+) or losses (–). Numbers shown in circles are the total number of genes present in the extant species or ancestral species (represented as nodes within the phylogeny) genome. Species abbreviations: Agam, *Anopheles gambiae* (mosquito); Amel, *Apis mellifera* (honeybee); Ath, *Arabidopsis thaliana*; Bmo, *Bombyx mori* (silkworm); Cel, *Caenorhabditis elegans*; Cneo, *Cryptococcus neoformans*; Ddi, *Dictyostelium discoideum* (slime mold); Dmel, *Drosophila melanogaster* (fruit fly); Ehi, *Entamoeba histolytica* (amoeba); Homo, *Homo sapiens* (human); Mus, *Mus musculus* (mouse); Osaj, *Oryza sativa* var. *japonica* (rice); Scer, *Saccharomyces cerevisiae*; Yli, *Yarrowia lipolytica*; Zfish: *Danio rerio* (zebrafish).

simply do not need a large and diverse number of fatty-acid containing compounds (in contrast to insects and plants), so only 1 or 2 genes are sufficient to synthesize all that is necessary.

It is difficult to say which of these possibilities is true without further knowledge of the FAR gene complement and associated functionalities from more species. Regardless, it is reasonable to assume that the genomic drift that produced the expansion of FAR genes in plants and insects is the underlying cause for their ability to synthesize and utilize a wide variety of fatty alcohol-based or derived compounds for a number of highly specialized functions.

Birth-and-Death Evolution and Selective Constraints: Histone Variant Diversification in the Germinal Cell Line

Multigene families often consist of structurally and functionally related genes that are usually clustered around specific genomic regions. The traditional view that a gene family producing a large amount of products needs to maintain homogeneity among its members [38] reinforced the notion that most multigene families were subject to concerted evolution, a process in which a mutation occurring in a repeat spreads all through the gene family members by recurrent unequal crossing-over or gene conversion. However, the increase in genomic molecular data during the last decade has revealed that most gene families encompass far too much genetic and functional diversity to be maintained by means of a homogenizing mechanism. Consequently, different alternative hypotheses have been put forward in order to account for the high diversity and functional differentiation exhibited by the members of different eukaryotic gene families. Among them, the birth-and-death model of evolution (which promotes genetic diversity) has often constituted the alternative hypothesis to concerted evolution [15].

Birth-and-Death Long-Term Evolution of Histone Multigene Families

In eukaryotes and some archaebacteria the members of the histone multigene families encode small basic proteins that are associated with the hereditary material in a nucleoprotein complex called chromatin, which allows for a high level of compaction of genomic DNA within the limited space of the nucleus and also provides the scaffolding upon which most DNA metabolic functions (i.e. replication, transcription and repair) take place. However, the different histone families display a high degree of heterogeneity among their members, depending on their structural and functional role in the nucleosome (the chromatin subunit) as well as depending upon whether the chromatin structure is in a somatic or a germinal setup. In addition, post-translational histone modifications also influence changes in chromatin structure both directly and indirectly by targeting or activating chromatin-remodeling complexes. Histone modifications intersect with cell signaling pathways to control

gene expression and can act combinatorially to enforce or reverse epigenetic marks in chromatin [39, 40].

Histones have been used (together with rDNA) to showcase archetypal examples of multigene families subject to concerted evolution during the last 4 decades. However, the notion of this mechanism representing the major long-term evolutionary mode of these proteins has been abandoned given the high diversity and functional differentiation exhibited by the members of the different histone families. On the contrary, it has now been clearly demonstrated that the long-term evolution of the histones can be better described by a birth-and-death model of evolution based on recurrent gene duplication events and strong purifying selection acting at the protein level (e.g. [16]). This mode of evolution eventually leads to the functional differentiation of new gene copies through a process of neofunctionalization or subfunctionalization [40].

Selective Constraints and Histone Diversification in Different Chromatin Setups

Eukaryotic DNA is packed into different chromatin configurations in somatic and germinal cells. Somatic chromatin is formed by the repetition of nucleosomes [41], each consisting of an octamer of core histones (2 of each H2A, H2B, H3 and H4) around which 2 left-handed super-helical turns of DNA (approximately 146 bp) are wrapped. The nucleosomes are joined together in the chromatin fiber by short stretches of linker DNA that interact with linker H1 histones, resulting in an additional folding of the chromatin fiber. Germinal chromatin displays a high degree of heterogeneity depending on sex (male or female) and taxonomic group. Thus, while a nucleosome-based chromatin organization is prevalent in the case of the female germinal cell line (i.e. oocytes), the extreme reduction in the size of the sperm nucleus has led to a drastic reorganization in the male-specific chromatin in which nucleosomes have been replaced by nucleoprotein structures able to produce a tighter packaging of DNA [40].

Sperm chromatin is unique in that most, if not all, is tightly heterochromatinized within the highly compacted sperm nuclei thanks to its association with sperm nuclear basic proteins (SNBPs) [42]. In contrast to the proteins of somatic chromatin (histones), SNBPs exhibit a greater compositional heterogeneity and can be grouped into 3 major types based on structural and compositional considerations. The first is the histone type (H-type) SNBPs, which are very similar to histones from somatic tissues and, therefore, produce a chromatin organization identical to that observed in somatic cell nuclei. The second type consists of protamines (P-type SNBPs), which constitute a group of heterogeneous, small, arginine-rich proteins that result in a tighter packaging of DNA within the sperm nucleus. The third type of SNBPs form a group known as the protamine-like proteins (PL-type), which are related to histone H1 and represent a structurally and functionally intermediate group between the H- and P-types [42]. The chromatin fibers resulting from the association of the different SNBP types with DNA all exhibit a fairly constant diameter in the range of 300–500 Å, independent of

the extent of protein folding of the SNBP type involved, which decreases from the H- to PL-type and from the PL- to the P-type [39].

Somatic chromatin is characterized by a nucleosome-based organization in which histones associate with each other and with DNA through different protein-protein interactions including those of an electrostatic nature. Histone proteins are thus subject to strong selective constraints in order to preserve their structure along with the nucleoprotein complex they form with DNA. However, the transition from somatic to germinal chromatin setups during spermiogenesis involves the replacement of histones by specialized SNBPs, leading to the progressive loss of a nucleosome-based chromatin configuration [39]. In this scenario, the functional constraints operating on histones in the germinal cell line are expected to be relaxed, allowing for a higher degree of variation within the different histone types (fig. 3).

Increased Birth-and-Death Histone Diversification in the Male Germinal Cell Line

Nucleosomes modulate accessibility of regulatory proteins to DNA and thus influence eukaryotic gene regulation. The evolution of chromatin remodeling mechanisms governing nucleosome organization at promoters, regulatory elements, and other functional regions in the genome unveil an interplay of sequence-based nucleosome preferences and non-nucleosomal factors in determining nucleosome organization within mammalian cells. The genetic diversity observed among histone family members bears critical implications for the structure and function of the nucleosome in different chromatin settings [43], involving the formation of H2A-H2B and H3-H4 dimers through different protein-protein interactions, including those of an electrostatic nature. When looking at the diversity within core histone families (fig. 3a), it seems that although one of each interacting partners is allowed to have a higher extent of variation (H2A and H3), the other maintains a conserved structure (H2B and H4). Molecular evolutionary studies carried out during the last 10 years have revealed that the long-term evolution of the histone H1 family, as well as of H2A, H3, and H4 core histone families, is governed by birth-and-death under a strong purifying selection acting at the protein level, in order to preserve a functional quaternary structure of the nucleosome core particle [40], able to efficiently bind and package the DNA, as well as to mediate different dynamic processes in chromatin metabolism [43].

However, information about the diversity and the evolution of H2B was lacking until very recently. The H2B family stands out among histones because of the low extent of diversification of its members (compared with H1, H2A, and H3 families) and the lack of specialized variants in the somatic cell lineage. Nevertheless, the H2B family is peculiar by displaying variants exclusively restricted to the male germinal cell lineage. For instance, 2 testis-specific variants have been described in humans so far, including TH2B (also referred to as hTSH2B) [44] and H2BFW (also known as H2BFWT) [45], both involved in the reorganization of chromatin during spermatogenesis. Furthermore, additional minor H2B variants with a lower extent of similarity

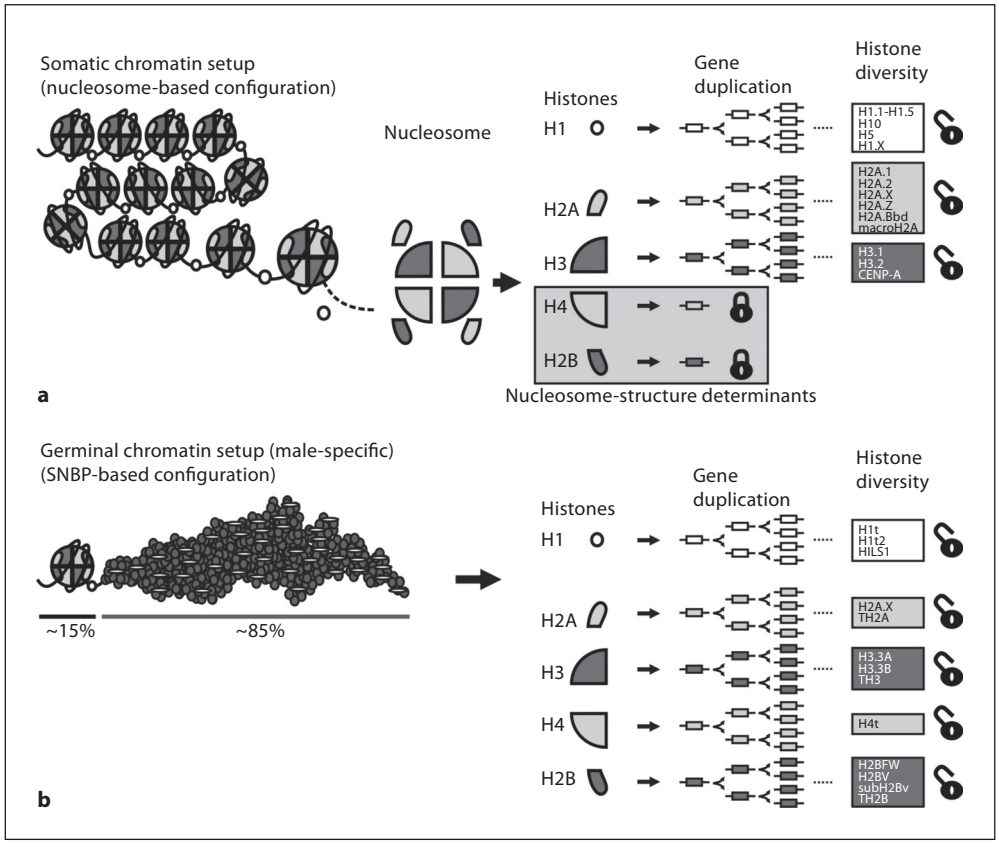


Fig. 3. Chromatin organization and histone diversification in the somatic and male-specific germinal cell line. **a** Histone H2B and H4 variant diversification is locked within a somatic chromatin setup, probably as a consequence of their essential role in maintaining the fundamental structural H2A-H2B and H3-H4 domains of the nucleosome core particle. In contrast, the variation presented by the H2A and H3 counterparts is responsible for imparting different functional and structural specificities to these domains, allowing for the specialization of local chromatin segments genome-wide. **b** The structural reorganization of chromatin during spermiogenesis leads to the loss of a nucleosome-based configuration in the male germinal cell line, lightening the evolutionary constraints operating on histone H2B and H4 evolution. Consequently the process of diversification within these histone families is unlocked allowing for the functional differentiation of germinal variants.

with canonical H2Bs have also been described in the male germinal line including subH2Bv, a sperm-specific histone identified in the bull *Bos taurus*; gH2B, a divergent H2B protein identified in *Lilium longiflorum*, involved in the packaging of chromatin in pollen; and H2BV, a variant first identified in *Trypanosoma brucei* that specifically dimerizes with H2A.Z. In addition, 2 novel H2B variants involved in pericentric heterochromatin reprogramming during mouse spermiogenesis, referred to as H2BL1 and H2BL2, have been recently identified [46], showing resemblance to subH2Bv and H2BFW, respectively.

The constraints driving the long-term evolution of the H2B family in the somatic cell line have been recently investigated, corroborating the presence of birth-and-death evolution under strong purifying selection, maintaining high levels of certain biased amino acids (lysine and alanine) which are important for the establishment of the correct interactions involved in the formation of the nucleosome [47]. On the other hand, and in contrast with other histones, H2B members are also subject to a very rapid process of diversification in the male germinal cell lineage (fig. 3b) involving the functional specialization of different histone variants, probably as a consequence of neofunctionalization and subfunctionalization events after gene duplication [47]. This is specifically evident in the case of the H2BFW variant that evolves almost at the same rate as the quickly evolving histone H2A.Bbd which is also involved in mammalian spermiogenesis [48].

The lack of diversity within the H2B and H4 families has been regarded to be the result of their essential role in the maintenance of the fundamental structural H2A-H2B and H3-H4 domains of the nucleosome. By contrast, the variation presented by the H2A and H3 counterparts would be responsible for imparting different functional and structural specificities to these domains [43]. Such a hypothesis would be consistent with the increase in H2B diversity observed in the male germinal cell line where a dramatic change in chromatin conformation takes place during spermiogenesis. Two conclusions can be drawn from this. First, H2B variation implicitly suggests the possibility of H4 variation. Indeed, the few H4 variants described to date are mostly circumscribed to the testis [49]. Second, the diversification of H2B and H4 histones would be absent from the female germinal cell line (i.e. in oocytes) due to the prevalence of a nucleosome chromatin organization, which would only be compatible with H1 variants such as H1oo and H1M/B4. It thus seems that the reorganization of chromatin structure during spermiogenesis might have affected the evolutionary constraints driving histone H2B evolution, leading to an increase in diversity. However, with the exception of a few structural studies [50], little is known about the specific role performed by the testis-specific H2B variants. Further studies will be needed in order to clearly decipher the connection between the relaxation of the evolutionary constraints described here and the drastic structural chromatin transitions involved in spermiogenesis.

Mixed Effects of Birth-and-Death and Concerted Evolution: the 5S rDNA Gene Family in Fishes and Molluscs

In eukaryotes, rDNA is generally arranged in 2 different gene clusters (multigene families), each composed of hundreds to thousands of gene copies. While the major cluster (45S rDNA) comprises the 18S, 5.8S, and 28S rRNA genes, the minor cluster (5S rDNA) comprises only 5S rRNA genes. The 5S rRNA gene consists of a transcriptional unit of ~120 bp, which is separated from the next unit by a non-transcribed

spacer (NTS). Although the 5S rRNA gene is highly conserved, the NTSs are variable both in length and in sequence [51]. Given the apparent homogeneity observed among the different copies, 5S genes have been used to showcase the archetypal example of a gene family subject to concerted evolution. However, the theoretical expectations made by this model are challenged by 3 major molecular evolutionary features displayed by the 5S rDNA family. First, several 5S gene variants have been found, constituting a dual system. Second, 5S rDNA divergent pseudogenes have been found in unrelated taxa. Third, the existence of different types of repeat units has been also corroborated based on the study of spacers. Consequently, different authors have proposed that the variation observed among 5S rDNA members best fits to a birth-and-death model of long-term evolution promoting genetic diversity [6].

Concerted Evolution of 5S rRNA Genes

Concerted evolution has been recently discarded (in favor of a birth-and-death mechanism) as the major model guiding the long-term evolution of several multigene families [6]. However, the case of rDNA seems to be otherwise more complex. Among animals, molluscs and fishes stand out for being the most widely studied groups of organisms with respect to 5S rRNA genes, displaying intense genetic dynamics. Studies on the 5S rDNA from oyster (genus *Crassostrea*) have revealed the existence of (1) two different genes (instead of one, as in the case of the major genes) encoding the minor 5S subunit, and (2) the localization of 5S rRNA genes in 2 pairs of chromosomes different from the chromosome pair (pair 10) where the major genes are located [52]. However, only 1 type of 5S rDNA tandem repeat was found in *Crassostrea* representatives. These results, together with the identification of a microsatellite at the 3′ end of 5S genes (potentially involved in the maintenance of tandem arrays), support the concerted evolution of 5S rRNA genes in these organisms.

Evidence supporting the concerted evolution of 5S rRNA genes has been also found in different fish representatives. For instance, decreased levels of intra- and interspecies nucleotide variation have been recently revealed in the 5S coding regions from fish species belonging to the family Moronidae [53]. Similarly to the case of oyster, the presence of microsatelllite sequences has been also identified at NTS regions. Different authors have suggested that the presence of short microsatellite sequences favors the maintenance of tandem arrays in multigene families. These sequences would act as 'hot spots' for recombination, facilitating gene conversion or unequal crossing-over and therefore, concerted evolution [54, 55].

Mixed Effects of Birth-and-Death and Concerted Evolution

Within molluscs, mussels also attract special interest due to the heterogeneity they display in 5S rDNA organization, including different types of repeat units with divergent spacers such as those identified in *Mytilus* species [56]. Recent studies on *Mytilus* species provided evidence for an apparent absence of interspecies differentiation across 5S coding regions, a notion reinforced by: (1) the lack of fixed differences

between species, and (2) the low levels of nucleotide variation found within 5S coding regions in comparisons between different types of units, suggesting the presence of independent evolutionary pathways leading to their differentiation [57]. Although these results do not fit the predictions made by the concerted evolution model, they can be still reconciled with a critical role for this evolutionary model in 5S rDNA evolution. Different studies have put forward a hypothesis in which the homogenization of rDNA units would occur locally within arrays, implying that selective mechanisms operate in the coding region, eliminating mutations without affecting spacer regions [58]. It is thus possible that a first stage of 5S rDNA evolution would had involved the generation of genetic diversity through recurrent gene duplications (birth-and-death), followed by the transposition of several units to different chromosomal locations, leading to the their subsequent independent concerted evolution. However, even though the observed patterns of 5S rDNA evolution could also result from a process of gene duplication and selection without invoking homogenization, a substantial effect of concerted evolution cannot be ruled out until the presence of heterogeneous selective constraints acting on different 5S types is demonstrated [57].

Many studies focused on the molecular organization and evolution of 5S rRNA genes have described the presence of 2 types of 5S rDNA units, especially in the case of fishes [59, 60]. The main difference between these sequences is essentially circumscribed to length polymorphisms in the NTS region, although variation in coding regions is sometimes observed, suggesting that the two 5S rDNA loci evolve independently. However, some reports suggest that both 5S rDNA types are not located in independent clusters, since different 5S variants have been found on the same PCR product displaying a tandem organization [61]. It thus appears that the existence of 2 types of 5S rDNA units constitutes a common trend in fish species [53, 55, 60]. This organization has been commonly referred to as 'dual expression system', where one type is expressed in both the somatic and the germinal (oocyte) cell line, while the other type is specific to oocyte cells. The presence of 5S rDNA units containing divergent types of NTSs was identified in the flatfish *Solea senegalensis*. Furthermore, a repeat unit containing the 5S rRNA gene linked simultaneously to 3 different small nuclear RNA genes (U1, U2, and U5) was described for the first time in this species (U2 snRNA appeared also in the NTS of the oyster *Crassostrea* [54]), probably representing pseudogenes [62]. Sequence divergence among tandemly arranged 5S rRNA and NTS sequences indicates that the rate of concerted evolution is insufficient to homogenize the entire array. Similar results have been described in stingrays [60], a coregonid fish, for which a significant amount of variation was reported in the 5S rRNA coding region and NTS sequences [63], as well as in species belonging to the genus *Brycon*, displaying high levels of divergence in the NTS region [5].

Birth-and-Death Evolution in Dual 5S rDNA Gene Systems

Several species of the family Batrachoididae have traditionally been used as model organisms within teleost fishes. For our studies, we have chosen 4 Venezuelan

species *(Amphichthys cryptocentrus, Batrachoides manglae, Porichthys plectrodon, Thalassophryne maculosa)* and the only European species within this family, the toadfish *Halobatrachus didactylus.* Two types of 5S rDNA units were found in *H. didactylus* and, given the lack of similarity between their NTS sequences, they probably do not share a common ancestral sequence. Although both types seem to represent functional genes, it cannot be concluded that a dual system of 5S rDNA is generally established in the Batrachoididae family since species displaying only one 5S rDNA type have also been found [55]. Given that the sequences of both coding regions and the NTSs are quite conserved in *H. didactylus*, concerted evolution seems to represent the more feasible model for this multigene family.

Although concerted evolution has been traditionally proposed to guide the long-term evolution of 5S rRNA genes, the birth-and-death model of evolution has been recently invoked in order to explain several cases in which homogenization is not observed [60, 64]. Under a birth-and-death model of evolution, 5S rDNA genes would be expected to display divergent variants in the genome, between-species clustering pattern in the phylogenies as well as the presence of pseudogenes. Genome rearrangements (e.g. gene duplications, deletions, insertions) are likely to have been involved in the evolution of 5S rRNA genes in the family Batrachoididae. The results of our analysis suggest that the 5S rRNA genes of the 4 species studied (and also of the European one) are derived from a dual 5S rDNA gene system which was already present in the genome of their common ancestor. However, while *A. cryptocentrus* and *B. manglae* have retained both types of 5S rDNA units, we have found only 1 type in *P. plectrodon* and *T. maculosa*. In these last 2 species, as well as in *B. manglae*, homogenizing mechanisms like those proposed by the concerted evolution model appear to have occurred. While *P. plectrodon* seems to have suffered a recent deletion event (and concerted evolution has not had enough time to act), one of the 5S rDNA types from *A. cryptocentrus* has undergone a higher degree of diversification. Therefore, the emergence of new 5S rDNA variants in *A. cryptocentrus* could be explained by birth-and-death evolution, and these variants could be maintained by purifying selection. Notwithstanding, we cannot exclude the possibility of some homogenization mechanisms reducing sequence divergence within each 5S rDNA unit in this species [61].

The birth-and-death evolution of 5S rDNA in fish species is also supported by the presence of pseudogenes, although the emergence of duplicated pseudogenes can also be explained by unequal crossing-over, one of the main mechanisms acting in concerted evolution [5]. In addition, NTS regions of *A. cryptocentrus* and *B. manglae* display a variable number of $(TG)_n$ or $(AG)_n$ microsatellites which could represent 'hot spots' playing an important role in homogenizing tandem arrays [54]. Furthermore, homogenization resulting from unequal crossing-over or gene conversion during concerted evolution would occur most frequently in regions of chromosomes closer to the telomeres [6]. In this regard, FISH studies using 5S rDNA probes have shown that minor ribosomal genes of *A. cryptocentrus* are located in a

subcentromeric position [55], which could hinder the action of the mechanisms that govern concerted evolution.

Birth-and-death has been proposed as a very important mechanism in guiding the long-term evolution of the 5S rDNA family in different organisms. Our results suggest that in many groups of molluscs and fishes the long-term evolution of 5S rRNA genes is most likely mediated by a mixed mechanism in which the generation of genetic diversity is achieved through birth-and-death (recurrent gene duplication), followed by the local homogenization of the different units through concerted evolution (probably after their physical transposition to independent chromosomal locations). In addition, it is important to bear in mind that to completely discern between the relative contributions of concerted evolution and birth-and-death evolution to the overall long-term evolution of 5S rRNA genes, it would be necessary to gather information on the complete set of 5S rRNA genes in different genomes. Although this has not yet been achieved for most 'higher' eukaryotes (including molluscs), it is not the case for certain groups of 'lower' eukaryotes. For example, in a complete genome study of 4 species of fungi, it was shown that the birth-and-death model without contribution of concerted evolution best characterizes the long-term evolution of 5S genes in those organisms [18]. In this case, the apparent homogenization among copies results from a combination of (1) recent gene duplication due to a gene duplication and insertion process similar to retroposon amplification and (2) rapid gene turnover derived from a high frequency of duplication/amplification events. Without a precise knowledge of the complete genome complement of these taxa and the subsequent comparison among closely related species, it is easy to misinterpret that homogenizing forces might also have an important role in the 5S gene evolution of those particular organisms.

Concluding Remarks

Over the long term, the birth-and-death process might result in a large variation in the number of genes or in the number of orthologous copies that would be visualized as gene family expansions (or contractions). The family size, therefore, would result from a trade-off between the stochastic birth-and-death process and the maintenance of genes required for proper function, as depicted by the case of the chemoreceptor system. Hence, the dynamic birth-and-death process has important evolutionary and adaptive implications: both gene gains and losses constitute a significant source of variation for evolutionary change. Indeed, DNA changes (in a particular duplicate) affecting the sensitivity or specificity in the detection of pheromones or related substances as food may be advantageous and might be fostered by shifts in ecological interactions. In so far, as the relevance of gene gains and losses to overall multigene family evolution is concerned, genomic drift plays a clear role in driving the divergence of entire multigene families. As we have shown in the case of the chemoreceptor

families, genomic drift can alter the composition of genes within a genome as well as between different species' genomes, encompassing an adaptive value behind these changes. Similarly, drift may have played a part in the case of the FAR gene family in facilitating the ecological adaptation of plants and insects to their environments through the ability to generate a range of fatty alcohols utilized for a variety of physiological purposes. Thus, genomic drift can be viewed as a driving force for evolving evolutionary novelty that can be exploited by a species as means for adaptation to various selective challenges.

Once selection starts operating over a multigene family, changes or shifts in selective constraints will affect the functional dynamics of the birth-and-death process. This mechanism is best exemplified by histone multigene families, where the relaxation of the selective constraints results in higher rates of functional diversification across family members which otherwise must be conserved in order to preserve the nucleosome-based structure of somatic chromatin. However, given the evolutionary patterns observed across 5S rDNA gene family members, an important effect of concerted evolution cannot be ruled out until the presence of heterogeneous selective constraints acting on different 5S types is demonstrated.

Over the last 2 decades many multigene families have been identified that undergo birth-and-death evolution, including former archetypal examples of concerted evolution, such as histones and rRNA genes. Far from the old controversies on the mechanisms driving the evolution of multigene families, the continuous stream of genomic molecular data keeps on creating an increasingly complex canvas of gene families filled with countless evolutionary nuances. In such a complex scenario, the birth-and-death model of evolution provides a framework to understand how multigene families originate and diversify, representing the principal mechanism guiding the long-term evolution of multigene families.

Acknowledgements

This work was supported by grants from the Xunta de Galicia (10-PXIB-103-077-PR to J.M.E.-L.), from the Ministerio de Ciencia e Innovación of Spain-MICINN (CGL2011-24812 to J.M.E.-L., and BFU2010-15484 to J.R.), and from the Junta de Andalucía and CeiA3 (Campus de Excelencia Internacional Agroalimentario to L.R., group BIO-219). J.M.E.-L. was supported by a contract within the Ramon y Cajal Subprogramme (Ministerio de Ciencia e Innovación of Spain-MICINN), and J.R. was partially supported by ICREA Academia (Generalitat de Catalunya).

References

1 Ohno S: Evolution by Gene Duplication. Berlin, Springer-Verlag, 1970.

2 Yang Z: Computational Molecular Evolution. Oxford, Oxford University Press, 2006.

3 Demuth JP, Hahn MW: The life and death of gene families. Bioessays 2009;31:29–39.

4 Gabaldon T: Large-scale assignment of orthology: back to phylogenetics? Genome Biol 2008;9:235.

5 Martins C, Wasko AP: Organization and evolution of 5S ribosomal DNA in the fish genome; in Williams CL (ed): Focus on Genome Research. Hauppauge, Nova Science Publishers, 2004, pp 335–363.
6 Nei M, Rooney AP: Concerted and birth-and-death evolution in multigene families. Annu Rev Genet 2005;39:121–152.
7 Lynch M: The Origins of Genome Architecture. Sunderland, MA, Sinauer Associates, 2007.
8 Hahn MW: Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol 2007;8:R141.
9 Csuros M: Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 2010;26:1910–1912.
10 Iwasaki W, Takagi T: Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics 2007;23:i230–239.
11 Vernot B, Stolzer M, Goldman A, Durand D: Reconciliation with non-binary species trees. Comput Syst Bioinformatics Conf 2007;6:441–452.
12 Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, et al: Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 2005;21:2596–2603.
13 Vieira FG, Sanchez-Gracia A, Rozas J: Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. Genome Biol 2007;8:R235.
14 Ingram VM: Gene evolution and the haemoglobins. Nature 1961;189:704–708.
15 Nei M, Hughes AL: Balanced polymorphism and evolution by the birth-and-death process in the MHC loci; in Tsuji K, Aizawa M, Sasazuki T (eds): 11th Histocompatibility Workshop and Conference. Oxford, Oxford University Press, 1992, pp 27–38.
16 Rooney AP, Piontkivska H, Nei M: Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. Mol Biol Evol 2002;19:68–75.
17 Rooney AP: Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. Mol Biol Evol 2004;21:1704–1711.
18 Rooney AP, Ward TJ: Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proc Natl Acad Sci USA 2005;102:5084–5089.
19 Zhang J, Dyer KD, Rosenberg HF: Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. Proc Natl Acad Sci USA 2000;97:4701–4706.
20 Vieira FG, Rozas J: Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol 2011;3:476–490.
21 Rooney AP, Ward TJ: Birth-and-death evolution of the internalin multigene family in *Listeria*. Gene 2008;427:124–128.
22 Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al: Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 2007;450:203–218.
23 Sanchez-Gracia A, Vieira FG, Rozas J: Molecular evolution of the major chemosensory gene families in insects. Heredity 2009;103:208–216.
24 De Bie T, Cristianini N, Demuth JP, Hahn MW: CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006;22:1269–1271.
25 Hahn MW, Han MV, Han SG: Gene family evolution across 12 *Drosophila* genomes. PLoS Genet 2007;3:e197.
26 Nei M: The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA 2007;104:12235–12242.
27 Long EO, Dawid IB: Repeated genes in eukaryotes. Annu Rev Biochem 1980;49:727–764.
28 Nei M, Niimura Y, Nozawa M: The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet 2008;9:951–963.
29 Nam J, Nei M: Evolutionary change of the numbers of homeobox genes in bilateral animals. Mol Biol Evol 2005;22:2386–2394.
30 Durnad D, Halldórsson BV, Vernot B: A hybrid micro–macroevolutionary approach to gene tree reconstruction. J Comput Biol 2006;13:320–335.
31 Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 1979;28:132–163.
32 Page R: Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. Syst Zool 1994;43:58–77.
33 Page R, Charleston M: From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 1997;7:231–240.
34 Antony B, Fuji T, Moto K, Matsumoto S, Fukuzawa M, et al: Pheromone-gland-specific fatty-acyl reductase in the adzuki bean borer, *Ostrinia scapulalis* (Lepidoptera: Crambidae). Insect Biochem Mol Biol 2009;39:90–95.

35 Moto K, Yoshiga T, Yamamoto M, Takahashi S, Okano K, et al: Pheromone gland-specific fatty-acyl reductase of the silkmoth, *Bombyx mori*. Proc Natl Acad Sci USA 2003;100:9156–9161.

36 Miwa T: Jojoba oil wax esters and derived fatty acids and alcohols: gas chromatographic analyses. J Am Oil Chem Soc 1971;48:259–264.

37 Rowland O, Zheng H, Hepworth SR, Lam P, Jetter R, et al: *CER4* encodes an alcohol-forming fatty acyl-coenzyme A reductase involved in cuticular wax production in *Arabidopsis*. Plant Physiol 2006; 142:866–877.

38 Thatcher TH, Gorovsky MA: Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. Nucleic Acids Res 1994;22:174–179.

39 Eirín-López JM, Ausió J: Origin and evolution of chromosomal sperm proteins. Bioessays 2009;31: 1062–1070.

40 Eirín-López JM, González-Romero R, Dryhurst D, Méndez J, Ausió J: Long-term evolution of histone families: old notions and new insights into their diversification mechanisms across eukaryotes; in Pontarotti P (ed): Evolutionary Biology: Concept, Modeling, and Application. Berlin, Springer-Verlag, 2009, pp 139–162.

41 Zlatanova J, Bishop TC, Victor JM, Jackson V, van Holde K: The nucleosome family: dynamic and growing. Structure 2009;17:160–171.

42 Eirín-López JM, Frehlick LJ, Ausió J: Protamines, in the footsteps of linker histone evolution. J Biol Chem 2006;281:1–4.

43 Talbert PB, Henikoff S: Histone variants – ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol 2010;11:264–275.

44 Zalensky AO, Siino JS, Gineitis AA, Zalenskaya IA, Tomilin NV, et al: Human testis/sperm-specific histone H2B (hTSH2B). Molecular cloning and characterization. J Biol Chem 2002;277:43474–43480.

45 Churikov D, Siino J, Svetlova M, Zhang K, Gineitis A, et al: Novel human testis-specific histone H2B encoded by the interrupted gene on the X chromosome. Genomics 2004;84:745–756.

46 Govin J, Escoffier E, Rousseaux S, Kuhn L, Ferro M, et al: Pericentric heterochromatin reprogramming by new histone variants during mouse spermiogenesis. J Cell Biol 2007;176:283–294.

47 González-Romero R, Rivera-Casas C, Ausió J, Méndez J, Eirín-López JM: Birth-and-death long-term evolution promotes histone H2B variant diversification in the male germinal cell line. Mol Biol Evol 2010;27:1802–1812.

48 Eirín-López JM, Ishibashi T, Ausió J: H2A.Bbd: a quickly evolving hypervariable mammalian histone that destabilizes nucleosomes in an acetylation-independent way. FASEB J 2008;22:316–326.

49 Wolfe SA, Grimes SR: Protein-DNA interactions within the rat histone H4t promoter. J Biol Chem 1991;266:6637–6643.

50 Li A, Maffey AH, Abbott WD, Conde e Silva N, Prunell A, et al: Characterization of nucleosomes consisting of the human testis/sperm-specific histone H2B variant (hTSH2B). Biochemistry 2005;44: 2529–2535.

51 Campo D, Machado-Schiaffino G, Horreo JL, Garcia-Vazquez E: Molecular organization and evolution of 5S rDNA in the genus *Merluccius* and their phylogenetic implications. J Mol Evol 2009;68:208–216.

52 Cross I, Vega L, Rebordinos L: Nucleolar organizing regions in *Crassostrea angulata*: chromosomal location and polymorphism. Genetica 2003;119:65–74.

53 Merlo MA, Cross I, Chairi H, Manchado M, Rebordinos L: Analysis of three multigene families as useful tools in species characterization of two closely-related species, *Dicentrarchus labrax, Dicentrarchus punctatus* and their hybrids. Genes Genet Syst 2010;85:341–349.

54 Cross I, Rebordinos L: 5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas* oysters: Does the (CT)n(GA)n microsatellite stabilize this novel linkage of large tandem arrays? Genome 2005;48:1116–1119.

55 Ubeda-Manzanaro M, Merlo MA, Palazon JL, Sarasquete C, Rebordinos L: Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in species of the family Batrachoididae. Genome 2010;53:723–730.

56 Insua A, Freire R, Ríos J, Méndez J: The 5S rDNA of mussels *Mytilus galloprovincialis* and *M. edulis*: sequence variation and chromosomal location. Chromosome Res 2001;9:495–505.

57 Freire R, Arias A, Insua A, Méndez J, Eirin-Lopez JM: Evolutionary dynamics of the 5S rDNA gene family in the mussel *Mytilus*: mixed effects of birth-and-death and concerted evolution. J Mol Evol 2010;70:413–426.

58 Kellogg EA, Appels R: Intraspecific and interspecific variation in 5S RNA genes are decoupled in diploid wheat relatives. Genetics 1995;140:325–343.

59 Pinhal D, Araki CS, Gadig OB, Martins C: Molecular organization of 5S rDNA in sharks of the genus *Rhizoprionodon*: insights into the evolutionary dynamics of 5S rDNA in vertebrate genomes. Genet Res (Camb) 2009;91:61–72.

60 Pinhal D, Yoshimura TS, Araki CS, Martins C: The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: an example from freshwater stingrays. BMC Evol Biol 2011;11:151.

61 Robles F, de la Herran R, Ludwig A, Rejon CR, Rejon MR, et al: Genomic organization and evolution of the 5S ribosomal DNA in the ancient fish sturgeon. Genome 2005;48:18–28.
62 Manchado M, Zuasti E, Cross I, Merlo A, Infante C, et al: Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: A new linkage to the U1, U2, and U5 small nuclear RNA genes. Genome 2006;49:79–86.
63 Sajdak SL, Reed KM, Phillips RB: Intraindividual and interspecies variation in the 5S rDNA of coregonid fish. J Mol Evol 1998;46:680–688.
64 Lopez-Piñon MJ, Freire R, Insua A, Mendez J: Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in some scallops (Bivalvia: Pectinidae). Hereditas 2008;145:9–19.
65 Tamura K, Dudley J, Nei M, Kumar S: MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 2007;24:1596–1599.
66 Hedges SB, Kumar S: The Time Tree of Life. New York, Oxford University Press, 2009.

José M. Eirín-López
Departamento de Biología Celular y Molecular
Universidade da Coruña, Facultade de Ciencias
Campus de A Zapateira s/n, ES–15071 A Coruña (Spain)
Tel. +34 981 167 000 (2257), E-Mail jeirin@udc.es, http://chromevol.udc.es