The life and death of gene families

Jeffery P. Demuth,[1]* *and Matthew W. Hahn*[2]

[1]Department of Biology, University of Texas, Arlington, TX, USA
[2]Department of Biology and School of Informatics, Indiana University, Bloomington, IN, USA

One of the unique insights provided by the growing number of fully sequenced genomes is the pervasiveness of gene duplication and gene loss. Indeed, several metrics now suggest that rates of gene birth and death *per* gene are only 10–40% lower than nucleotide substitutions *per* site, and that *per* nucleotide, the consequent lineage-specific expansion and contraction of gene families may play at least as large a role in adaptation as changes in orthologous sequences. While gene family evolution is pervasive, it may be especially important in our own evolution since it appears that the ''revolving door'' of gene duplication and loss has undergone multiple accelerations in the lineage leading to humans. In this paper, we review current understanding of gene family evolution including: methods for inferring copy number change, evidence for adaptive expansion and adaptive contraction of gene families, the origins of new families and deaths of previously established ones, and finally we conclude with a perspective on challenges and promising directions for future research.

Keywords: gene copy number; gene duplication; gene family; gene loss

Introduction

Almost 40 years ago, Susumu Ohno promoted the idea that gene and genome duplications are the principle creative force in evolution.(1) While many have agreed with the potential evolutionary importance of duplications since then, until recently the evidence for copy number changes remained confined to a limited number of examples. Over the last decade, comparative analyses of an explosively growing number of fully sequenced genomes demonstrate that the size and complement of gene families are even more dynamic than expected by most. These findings not only support Ohno's view, but complete sequence coverage also provides evidence for the evolutionary importance of many gene losses that could not be appreciated prior to the genomic era.

Abbreviations: BDR, birth and death rate; CNV, copy number variation; HGT, horizontal gene transfer; WGD, whole genome duplication; OR, olfactory receptor.

***Correspondence to:** J. Demuth, Department of Biology, University of Texas, Arlington, TX 76019, USA.
E-mail: jpdemuth@uta.edu

The impact of gene family evolution may have been particularly important to human evolution as the rate of gene gain and loss appears to accelerate(2) while nucleotide substitution rates have declined.(3,4) Per nucleotide, gene copy number changes may explain just as much of the genetic divergence between humans and chimpanzee as orthologous nucleotide substitutions.(5,6) The rapid rate of divergence in copy number evident from comparative genomics among species is also consistent with studies of copy number variation (CNV) within humans, where hundreds of CNVs are found.(7) On a per nucleotide basis, copy number variants represent a larger pool of variation available to selection than single nucleotide polymorphisms.(7) Furthermore, the rapid pace of copy number change suggests that natural selection has often acted on gene family size and may be at least as important to organismal differences, and particularly adaptation, as changes to protein coding or regulatory regions.

In the following, we review efforts to understand the magnitude, rate, and distribution of changes in gene family size, as well as the evolutionary forces governing these changes. First, we briefly discuss how gene families are defined operationally and the computational methods for counting gene gains and losses in a comparative genomics framework. Second, we review estimates of the rate of gene birth and death as well as the experimental and comparative evidence for adaptive expansions and contractions in the life of gene families. Third, we discuss mechanisms resulting in the origin of new families and the circumstances that result in gene family death. Finally, we end with a look at the obstacles to improving our understanding of gene family evolution and present what we think will be important focuses of future research.

What is a gene family?

Gene families are groups of genes descended from a common ancestor that retain similar sequences and often similar functions.(8) The concept of a gene family applies to both genes within a single genome (paralogs) and related genes between genomes (orthologs and paralogs),(8) though the former was the only definition used for a long period of time. The ability to compare the number of gene copies among species *via* comparative genomics has had the effect

of re-introducing a wider meaning of gene family to include both paralogs within a species and orthologous or paralogous genes between species. This wider meaning implies that every gene must belong to a gene family, even single-copy genes—otherwise there would be no sense in comparing the size of gene families among species if even one of the species had only a single copy. One can still find both the *sensu stricto* and *sensu lato* meanings of gene families in the literature today, even in the same paper (*e.g.*,[9]).

While Muller's dictum, ''every gene from a pre-existing gene'' [10] implies that all genes are ultimately descendants of one ancient progenitor and consequently belong to a single ancient gene family, there are at least two benefits of subdividing family membership based on sequence similarity: (I) because sequence similarity confers structural similarity[11] which in many cases confers a degree of functional similarity,[12] well annotated genes can be used to assign functions to lesser known genes with similar sequences and (II) comparison of gene family content across species may provide insight into the evolutionary pressures that have shaped adaptation and diversity.[13] For these reasons, gene families are often defined by clustering genes across species by sequence similarity.

The process of clustering genes into families is analogous to reconstructing organismal phylogenies and poses many similar difficulties.[8,14] For instance, both gene and species lineages evolve and proliferate at different rates. And just as not all of an organism's traits reflect the species phylogeny, a typical eukaryotic gene is comprised of a mosaic of functional domains that may reflect different evolutionary relationships.[15–17] These characteristics often make it difficult to determine the appropriate threshold for defining both higher taxa and gene families. This means that even simple results—such as the proportion of genes in a genome with a duplicate—will depend on the threshold used for clustering and may therefore differ for non-biological reasons from study to study. For gene family clustering, differential rates of domain sequence divergence within and among lineages is particularly problematic because ''hybrid genes'' are know to arise *via* chimeric fusions between partially duplicated genes (*e.g.*,[18,19]).

Methodologies for gene family clustering and phylogenetic reconstruction have also followed similar maturation processes. Methods for phylogenetic inference have shifted from phenetic clustering, to parsimony and likelihood methods that are better able to account for rate variation among lineages and data types.[20] Similarly, the hierarchical clustering of proteins into superfamilies, families, subfamilies, and "Atlas entries'' based on thresholds of pairwise amino acid similarity[8] has given-way to more elaborate methods that try to overcome some of the problems posed by multi-domain proteins (*e.g.*,[21]). The variety of methods for detecting homology among protein sequences, fall along a continuum from complete automation to extensive manual supervision and post-clustering database curation. At the coarsest scale, a number of studies suggest that 40% amino acid similarity is a minimum to make inferences of functional similarity,[22] but variation among clustering methods (and thresholds within methods) has a pervasive impact on the absolute numbers and membership of genes in particular families.[23,24] The relative composition of gene families across taxa is typically less affected by method;[25] consequently, comparative analyses are usually robust within a study, but direct comparisons across studies can be problematic. Clustering methods and their corresponding databases of gene (or protein) family classification are comprehensively reviewed elsewhere,[26,27] so in the remainder of our review we limit discussion of clustering artifacts to those that directly impact inferences about gene family evolution.

Computational methods for measuring changes in gene copy number

Gene family changes result from differential duplication and loss of genes among evolutionary lineages. To understand the evolutionary forces governing this process it is first necessary to gain an accurate accounting of the number of gains and losses in any particular lineage. Given the combined difficulty of defining families and the heterogeneous quality of genome annotation, such an accounting is not a trivial undertaking. Beyond simple pairwise comparisons between genomes, two computational methods have been employed for this type of analysis. Thus far, the more widely adopted method compares a well-supported species tree to phylogenies for each gene family based on their nucleotide or protein sequences. By reconciling the gene-tree with the species-tree, one can infer the number of gene gains and losses on each branch of the species phylogeny.[28] There are two primary shortcomings of the tree reconciliation method. First, when gene-trees are not accurate there is a bias in the inferred pattern of duplications and losses. Specifically, inaccurate gene-trees cause the method to infer an excess of recent duplications and an excess of ancient losses.[29] Second, the tree reconciliation method does not provide a straightforward means to infer which evolutionary forces were responsible for the observed changes in the family size.

We previously developed a second strategy for inferring gene gains and losses that also provides a probabilistic framework for inferring evolutionary mechanisms. Our method uses maximum likelihood to infer family sizes at each internal node in the species phylogeny; the number of gene copies in each family and estimates of divergence time are the only data needed. The method simultaneously estimates the birth and death rate (BDR) that best fits the distribution of observed changes for all gene families.[30] An

advantage of our method is that by estimating the "average rates" of birth and death in genomes, we are able to make statistical inferences about the likelihood that any particular change in gene family size is the product of a purely stochastic process.[2,25,31,32] This likelihood method is also not without weaknesses. First, if multiple gains and losses occur on a particular branch the model only infers the net change (*i.e.*, always estimates the minimum number of changes); consequently, uncertainty in the number of changes grows with divergence time and very long branches will underestimate the true number of changes. Although this weakness precludes comparisons of anciently diverged taxa, it also yields a conservative estimate of the amount of change. Second, the model is currently constrained to equilibrium genome size (*i.e.*, on average, birth rate = death rate), so it is not useful for comparisons between taxa separated by whole genome duplications (WGDs). However, many closely related taxa maintain relatively constant gene numbers, suggesting that the equilibrium assumption will not be onerous for many comparisons, and comparisons between tree reconciliation and likelihood methods demonstrate very similar results despite these assumptions.[2,32] The final limitation of the likelihood method is that although qualitatively similar results are produced over a meaningful range of threshold values, the absolute values of change and rate estimates are sensitive to the details of gene family clustering method. Given the limitations of both tree reconciliation and the likelihood method, the best practice may be to use both in concert.[32]

Rates of gene family change

Rates of gene gain and loss are determined by an often difficult to disentangle interplay of mutation, fixation, and retention probabilities. Analyses of gene family evolution in yeast,[31] mammals,[2,25] and flies[32] find that genes appear to be gained and lost at remarkably similar rates (0.0020, 0.0016, and 0.0012 gains and losses/gene/my respectively).[25,31,32] Interestingly, improvements in the ability to model changes in BDR among lineages within a tree showed that while the average BDR of these anciently diverged groups is very similar, significant variation exists within groups. For example, in mammals the rate of gene turnover has nearly doubled in the primate lineage (0.0024 gains and losses/gene/my) compared to the lineages containing dog, mouse, and rat (0.0014 gains and losses/gene/my). A further acceleration has occurred in the great ape lineage (0.0039 gains and losses/gene/my) such that humans and chimps gain and lose genes almost 3X faster than other, non-primate, mammals.[2] Similar BDR variation occurs within the genus *Drosophila* (range from 0.0006 to 0.0193 gains and losses/gene/my); however, the degree of rate heterogeneity must be interpreted with caution since the depth of sequencing coverage is heterogeneous across the 12 species and the fastest rates are found on lineages with low coverage genomes.[32]

Estimates of the BDR in these groups are consistent with previous estimates based on the number of recent gene duplicates.[33,34] Using the number of paralogs with silent site divergence $\leq$1% as an estimate of the number of new duplicates before losses accrue, new duplicates appear to be "born" at the rate of 0.001–0.016 per gene per million years for a broad sample of eukaryotes.[33,34] A summary of BDR estimates based on this methodology suggest that the death rate of recent duplicates is at least an order of magnitude higher than the birth rate (Table 8.1 in [35]), at least partly because many young duplicates eventually become pseudogenes.

Adaptive expansion of gene families

Rapid gene family expansion in phenotypically important genes suggests scenarios wherein adaptive natural selection favors additional copies either for increased dosage or for an increased arsenal of molecular weaponry. To assess the role of natural selection in driving gene family expansion, let us consider the experimental and comparative genomics evidence for this phenomenon.

Direct evidence: gene amplification

There is substantial direct experimental evidence for an adaptive increase in gene family size in bacteria and some eukaryotes. In most studies of experimental evolution, rapid gene family expansion is clearly a product of selection to increase dosage. Typically referred to as "gene amplification," rapid accumulation of tandemly arrayed gene duplicates is often induced by an environmental stressor such as toxic or poor nutrient environments and mediated by transposable element activity. Bacterial gene amplification occurs in response to growth on non-standard media, and is a normal response to antibiotic exposure where extra gene copies promote increased metabolism of environmental constituents.[36] Human health may be negatively impacted by amplification of the cholera toxin gene region (*ctx*) and *Haemophilus influenzae* capsule formation genes as both are correlated with increased virulence of these human pathogens.[37,38] Although it remains controversial, gene amplification in *Escherichia coli* may also explain the phenomenon of "directed" or "adaptive mutation"[39] by increasing dosage of a mutant protein with limited functionality while simultaneously increasing the mutational target size for mutational reversion.[40,41] Upon reversion of one copy to full functionality, the remaining copies become superfluous and deletion-biased mutation rates or selection for replication efficiency

result in their loss. Thus, gene amplification is reversible in bacteria such that when the need for increased dosage is removed the genome reverts to the original copy number. This ''accordian'' of gene family expansion and contraction suggests that selection can fine tune gene dosage by adjusting gene copy number in organisms with very large population sizes.

In eukaryotes gene amplification appears to be less common. However, this may be due to the ineffectiveness of selection in small populations rather than actual differences in mutational input. The population size effect is reinforced by the fact that among eukaryotes, adaptive amplification appears most frequently in yeast, followed by insects, and is rare or absent in vertebrates. Gene amplifications in yeast are responsible for resistance to copper toxicity,[42] growth under resource limited conditions,[43] and dosage compensation for loss of one pair of histone genes (HTA1–HTB2).[44] In several insects, independent amplifications of certain esterase genes are responsible for resistance to organophosphate pesticides.[45,46] The most dramatic case of expansion is a 250-fold copy number increase in resistant strains of the mosquito *Culex pipiens*.[47] There is little evidence for an adaptive role of gene amplification in vertebrates, though mammalian cell lines provided an early example of gene multiplication in response to selection (*e.g.*,[48]). Gene amplification does occur in vertebrates, however, and is a principal pathology of some human cancers.[49] For example, the HER-2/neu oncogene is amplified up to 20-fold in some breast cancer tumors and copy number is a significant predictor of survival and time to relapse.[50] Although the occurrences of gene amplification cited above all concern increasing dosage of the same gene product, Francino[51] proposed that this process might also promote radiation of duplicates into new functions. The ''adaptive radiation'' model is consistent with the idea that selection for duplicative mutations *per se* will promote rapid gene family expansion, but the model's predictions overlap with those for other hypotheses requiring adaptive point mutations.[52]

Evidence from comparative genomics

Most cases of gene family expansion are evident only from a comparative analysis of copy number among extant lineages. Inferences concerning the adaptive significance of copy number expansions can be problematic because large families are expected to show large changes purely owing to their large size. Furthermore, evidence from nucleotide substitutions that suggests fixation of individual paralogs was driven by positive selection will probably be only a fraction of the families that have actually been selected for increased copy number. Many models of gene duplicate evolution may involve adaptive changes in copy number but never show evidence of positive selection at the nucleotide level.

We recently developed a method for computational analysis of gene family evolution (CAFE)[30,31] that provides the statistical machinery necessary to make probabilistic statements about whether the observed differences in gene family size among extant species are likely to be due to natural selection. To date, CAFE has been used to analyze gene family evolution in yeast,[31] flies, [32] and mammals.[2,25] In each group of taxa a number gene families (1.6–3%) have experienced sufficiently large changes in copy number to reject the null hypothesis of neutral evolution (at a false discovery rate <0.01). Among these families some functional categories evolve rapidly in all three groups: immune defense/stress response, metabolism, cell signaling, chemoreception, and reproduction related families (Table 1). Related functional categories (host evasion, metabolism, and environmental sensing) also constitute large lineage-specific expansions in many prokaryotic genomes.[53,54]

Table 1. Summary of gene family changes in three well sampled taxa. Because these studies use different clustering methods the absolute numbers of families are not comparable between taxa

Taxon	Families present in MRCA	Lineage-specific families	Whole family extinctions[a]	Rapidly evolving families[b]	Functions of rapidly evolving families
Mammals	9,990	2,278	1,421 (14.2%)	164 (1.6%)	Immune defense and response, transcription, translation, brain and neuron development, intercellular communication and transport, reproduction, metabolism, chemoreception
Drosophila	11,434	4,129	2,220 (19.4%)	343 (3%)	Defense response, proteolysis, trypsin activity, protein binding, response to chemical stimulus, and zinc ion binding
Yeast	3,517	NA	NA	56 (1.6%)	Stress response, metabolism, flocculation, myosin

Included genomes: Mammals (human, chimpanzee, mouse, rat, and dog); Drosophila (sechellia, simulans, yakuba, erecta, ananassae, persimilis, willistoni, mojavensis, virilis and grimshawi); Yeast (*S. baynus*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, *S. cerevisiae*). MRCA: most recent common ancestor.

[a]Independent extinctions of the same family are not counted separately here.

[b]Rapidly evolving families are defined by deviation from the rate expected by maximum likelihood estimation (see text for additional details). Based on data from Refs. [25,31,32].

The consistent pattern of expansion in gene families related to infection and disease suggests that co-evolution of immunity and virulence commonly involves reciprocal expansion of lineage-specific gene families in hosts and pathogens. Genes with these functions are also intriguing because they include some of the most rapidly evolving genes at the nucleotide level in *Drosophila* and mammals. For example, analysis of the *Drosophila* innate immune system, shows that pathogen recognition proteins evolve more often by positive selection on nucleotide changes, while proteins responsible for clearing infections (effector proteins) evolve more often by changes in copy number.[55] The pattern is less clear in mammals, where at least some expansions consist of natural selection favoring retention of duplicates that likely represented already divergent alleles (*e.g.*, MHC genes[56]) and/or adaptive divergence following duplication (*e.g.*, immunoglobulins[57]).

While the examples above are interesting because the same functions appear to evolve *via* gene family expansion in widely divergent organisms, the particular gene families that undergo expansion for the given functions are typically lineage-specific. Furthermore, expansion of some families clearly seems relevant to the organismal biology. For example, the *cathepsin B* family expansion in Aphids may play a role in their specialized high-sugar low-protein diet,[58] and the flocculin family expansion in *Saccharomyces cerevisiae* is probably the result of artificial selection during their domestication for brewing beer.[31] The *PRAME* family has undergone independent expansions in mouse, primates, and humans.[59] The normal expression of preferentially expressed antigen of melanoma (*PRAME*) genes is in testis, but they are also expressed in tumors, and have experienced considerable positive selection in the primate lineage.

To illustrate the lineage specificity of most gene family change, Fig. 1 shows the proportions of mammalian gene families that change size. Note that on each branch in the tree, the vast majority of gene family sizes remain static. In total, ~49% (4,893/9,990 families) of gene families change size, roughly half of which change on only one branch (56%; 2,754/4,893). This lineage specificity of change in families and functions implies that adaptation via copy number change is not a peculiarity of specific gene families: rather, it is a general mechanism that affects many different gene families depending on lineage-specific evolutionary pressures.

In addition to the potential for adaptive expansion of gene families, it has also been suggested that some gene families may be constrained in their ability to change size. This view is most often framed in terms of the ''balance hypothesis,'' which posits that genes that require more stoichiometric balance due to their interactions with other proteins are less likely to expand by single gene duplication (reviewed in Ref.[60]).

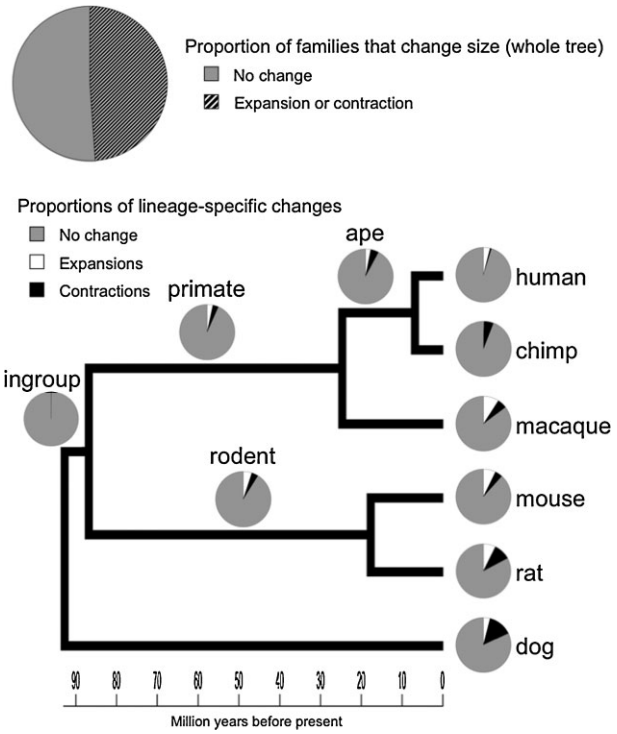

Figure 1. Distribution of changes in gene family size in mammals. The top chart shows the combined proportion of changes for the whole tree (proportion with expansion or contraction = # families that change size/total number of families in the mammalian MRCA). Charts along the mammalian tree show proportions of families in each lineage that gained (expansions) or lost (contractions) genes. The scale bar represents millions of years. (Redrawn from Ref.[25] with updated analysis from Ref.[2])

However, these families may experience greater expansion following WGD because in contrast to single gene duplications, WGD may maintain balance among dosage-sensitive partners. In support of the balance hypothesis, two analyses of duplicate retention following *Arabidopsis* WGD conclude that the functional classes of retained duplicates differ depending on the scale of the duplication.[61,62]

Adaptive contraction of gene families

While rapid gene family expansion may often be an indication of positive selection, the evolutionary pressures responsible for gene family contraction are less clear. Both neutral and adaptive explanations have been proposed. The evidence for adaptive gene loss remains rare and in most cases does not exclude the possibility that changes were the result of neutral processes. Indeed, the majority of evidence suggests that gene loss is more commonly the result of nonsense mutations drifting to fixation because they are not deleterious when natural selection is relaxed.

Direct evidence: silent genes

In principle, direct accounting for patterns of gene loss should be more straightforward than gaining experimental evidence for gene family expansions because silenced genes are typically not excised from the genome immediately. Therefore, quantifying pseudogene number provides a direct metric of the extent to which gene loss has impacted gene family sizes. This strategy has been fruitful in the study of many gene families that undergo rapid lineage-specific changes. For example the pseudogenization patterns of smell and taste receptors in mammals[63,64] and flies[65] suggests that the composition of chemoreceptors organisms maintain is specific to their diet and or habitat. This is perhaps no more striking than in human and other primate genomes where we have experienced extensive olfactory receptor (OR) family contractions resulting in our genomes containing hundreds of pseudogenes.[66] Paradoxically, the beginning of this massive loss of primate ORs may have coincided with the expansion of an opsin gene family that conferred trichromatic vision in Old World monkeys[67] (but see Ref.[68]). The remaining active copies of ORs show signatures of selection for sequence divergence in humans[69] as well as increased copy number in some human OR subfamilies.[25]

The above losses are most easily explained by changes in the environment causing a gene to no longer be essential to organismal fitness. Subsequently, nonsense mutations are able to drift to fixation because they are not sufficiently deleterious to be excluded by natural selection. It has also been proposed that in some cases gene loss may be adaptive, particularly in human evolution.[70] For instance, a number of gene losses have been attributed to this ''less is more'' hypothesis, including *MYH16*,[71] *CMAH*,[72] and *CASPASE12*.[73] In each case the null allele of these genes has been argued to confer a selective benefit (*MYH16*: capacity for brain case increase, *CMAH*: immune function and brain evolution, *CASPASE12*: protection from severe sepsis).

Evidence from comparative genomics

While the study of pseudogenes provides direct evidence for lineage specific contraction of some eukaryotic gene families, it is less useful for prokaryotes. In prokaryotes, deletion-biased mutation rapidly removes non-functional DNA and precludes the discovery of all but the most recent pseudogenes. Figure 2 illustrates the low proportion of pseudogenes in prokaryotic genomes relative eukaryotes.[74] Note that the relatively low proportion of pseudogenes in *Drosophila* may also be explained by deletion-biased mutation.[75] A consequence of the lack of pseudogene retention is that gene loss is best inferred by comparative methods (*e.g.*, missing orthologs in closely related taxa).

Although gene loss is a common theme in many organisms, it is most documented and most dramatic in obligate host-associated bacteria. Genomes of *Mycoplasma*, *Rickettsia*, *Chlamydia*, *Buchnera*, *Borrelia*, and their other parasitic and endosymbiotic relatives are among the smallest of all self-replicating organisms. Genome reduction in these bacteria often involves loss of hundreds to thousands of genes compared to closely related free-living taxa. The mechanism of this dramatic expulsion of genes is probably relaxed selection on genes that become superfluous after the bacteria adopt a parasitic (or symbiotic) lifestyle. The efficacy of purifying selection to retain these genes is further limited by reduced population size imposed by vertical transmission through the eukaryotic host, and because of deletion-biased mutation, pseudogenized and nearly neutral functional copies are removed from the genome (reviewed in Ref.[76]). Interestingly, the genes remaining after genome reduction are largely predictable based on models of the species' metabolism. For example the majority of the genes retained in the endosymbiotic bacteria *Buchnera* can be predicted from knowledge of its ancestral genome content and current physiology, suggesting that gene content is indeed shaped by natural selection.[77]

Massive gene losses are also found in at least two free-living bacteria, *Prochlorococcus* and *Pelagibacter*.[78,79] These marine bacteria are two of the most abundant organisms on earth and genome reduction is not easily explained by a model emphasizing relaxed selection as above. Instead, gene loss is suggested to be an adaptive response to more efficient replication. Because population sizes are sufficiently large the weak effect of a more ''streamlined'' genome may be enough to confer an evolutionary advantage and drive fixation in the population.[78,79]

Another group of gene families that consistently appear in lists of rapidly evolving families—but especially involve gene loss—are those involved in sensing the environment. As

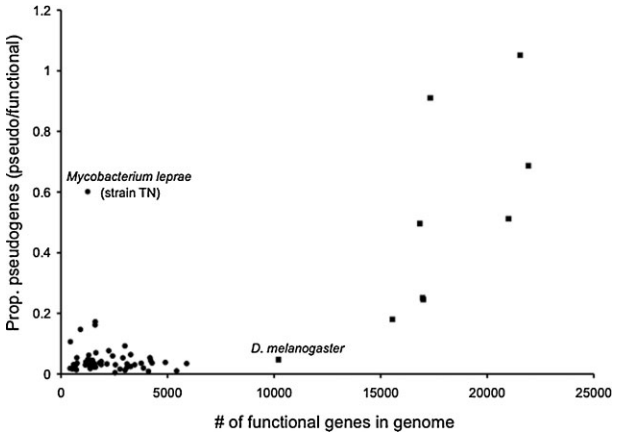

Figure 2. Proportion of pseudogenes relative to functional genes in prokaryotic (circles) and eukaryotic (squares) genomes. Data collected from pseudogenes.org[74] in March 2008.

BioEssays 31:29–39,

mentioned above, the largest superfamily of genes in mammals belong to the ORs, which confer our sense of smell. Among fully sequenced mammalian genomes, the number of functional OR genes ranges from 265 in Platypus to 1,207 in rat.[66] The OR superfamily has experienced a complex lineage specific history of expansions and contraction resulting in net gains in the mouse, rat, dog, and cow genomes, but large decreases in primates.[25,66,80] In *Drosophila*, chemoreceptor families with functional similarity to mammalian ORs have undergone rapid losses in multiple species in a pattern consistent with selection for specialization on certain host plants.[65,81]

The most abundant gene losses in eukaryotes occur following WGD. Lynch[35] summarized the number of genes retained following polyploidization in five animals, three plants and *S. cerevisiae*. The highest retention occurred in a frog (77%) and corn (72%), but all but one of the remaining lineages lost 50% or more of their genes. The largest reduction was in yeast where only 8% of duplicates from an ancient WGD have been retained.

The birth of gene families

Gene families do not only expand in size, they also expand in number (Table 1). New families typically originate with "orphan" genes[82,83] and can arise in three ways: (I) duplicate copies may become sufficiently divergent that they are no longer recognized as members of the same family[84] (II) genes can be horizontally transferred,[85–87] and (III) new genes can originate *de novo* from previously non-coding sequences.[88] In comparative studies, new families may also be incorrectly inferred due to complete loss of the gene family in related taxa. Clustering criteria have an important influence on the perceived number of gene family "creations" because tighter clustering results in more families that appear to have no relationship to other sequences. Most manually supervised clustering methods are biased against creating new families,[24] which consequently yields an estimate of novel families that is too low. Even for automated clustering, the choice of thresholds will impact the apparent number of new families (Fig. 3).

The effects of clustering criteria make inferring the pace of gene family origination problematic for whole genome analyses. This difficulty is exacerbated by heterogeneous sequence depth and/or annotation quality among taxa. For example, genomes are replete with species-specific single-gene families. In many cases these are *ab initio* gene predictions with no functional evidence, and are consequently likely to be artifacts of the annotation process. In other cases these apparent orphans may be artifacts of sparse taxon sampling, where increased sequencing of close relatives reveals orthologs in other species. In such cases the number of new orphans decreases as a function of taxon sampling, thus providing an estimate of the actual number.[89] These *bona fide* orphans constitute the origin of novel families.

Most lineage-specific families have only a small number of genes, but in a few cases families arise and undergo rapid expansion. For example, substantial human expansions have occurred in the primate specific FAM90A,[90] and *morpheus*[91] gene families, as well as mammal specific DUF1220 domain containing gene families[92,93] While additional examples of expanded lineage-specific gene families in other taxa exist (*e.g.*, nuclear receptors in nematodes[94]), the majority of these are examples of the duplication of pre-existing genes followed by neo-functionalization. Because their evolution remains constrained by vertical inheritance these "novel" families are unlikely to introduce radically new functions.

In contrast, horizontal gene transfer (HGT) potentially introduces novel gene families with far more foreign functions. HGT is most common in bacteria and archaea and is famously responsible for transferring both antibiotic resistance and increase virulence among unrelated human bacterial pathogens (reviewed in Ref.[95]). Gene transfers between bacteria and archaea have also conferred extreme evolutionary novelty across kingdoms. For example, transfer of genes from hyperthermophilc archaea to bacteria confer the ability to grow at $>80^{\circ}$C.[96] Eukaryotic HGT among nuclear genomes is very rare, although transfer from bacteria

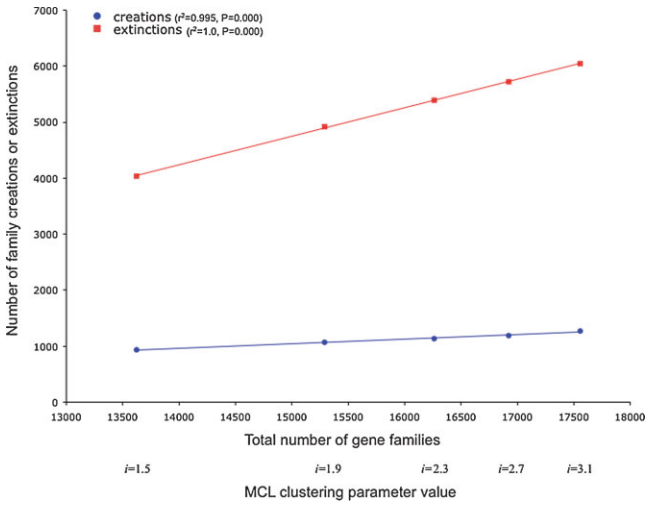

Figure 3. Effects of clustering threshold on the inferred numbers of newly created gene families (creations), whole family losses (extinctions), and total number of gene families. A Markov clustering method was used to cluster genes for human, chimpanzee, mouse, rat, and dog (redrawn from[25]). The inflation parameter (*i*) determines the stringency threshold for gene family inclusion. Higher values of *i* require genes to have higher sequence similarity and result in more, smaller families. higher thresholds also result in more inferred creations (blue) and extinctions (red). *r*-values represent the correlation between number of gene families and number of creations (or extinctions).

unicellular phagotrophic lineages is not uncommon;[97] and HGT from *Agrobacterium* to its plant hosts[98] and *Wolbachia* to its animal hosts[87] has also been documented. The rarity of nuclear gene transfer contrasts starkly with the permissiveness of plant mitochondria where HGT is relatively common.[99]

The final source of novel families comes from the *de novo* genesis of genes. While the gain of function from previously noncoding DNA is expected to be very rare, it has apparently given rise to several recently evolved testis-expressed *Drosophila* genes.[88] These genes yield very short transcripts, and it is unknown whether they encode proteins or functional RNAs. New genes may also arise by fusion of genes during a partial duplication.[19] In the case of the gene *jingwei* in *Drosophila*, a retro copy of an *alcohol dehydrogenase* gene captured exons of an unrelated gene and subsequently diverged in expression and function.[18]

The death of gene families

In some cases the loss of a gene will result in the extinction of an entire family. In contrast to the birth of new families, the death of gene families is the simple continuation of the loss of genes and does not involve any special evolutionary processes. Naively, we might expect that the complete loss of a biochemical function *via* loss of the last member of a gene family would typically be deleterious and consequently rare. Therefore, when gene family losses occur they may serve as indicators of shifts in the physiological constraints of an organism. For example, changes in diet affect the constraints imposed on different enzymatic pathways. Losses of genes in the *GAL* pathway result in an inability of some yeast species to metabolize galactose[100] and a variety of heterotrophic eukaryotes have lost the ability to synthesize nine amino acids which they are able to acquire from their diet.[101] More generally, genome wide analysis in animals suggests that gene families producing metabolic enzymes most frequently undergo independent extinction in multiple lineages.[23] This may indicate that shifts in nutrient availability or acquisition are most often responsible for conditions that permit gene family extinction.

There are a number of ways in which gene families can appear to go extinct: (I) deletion or pseudogenization of all members of a family, (II) accelerated protein evolution of individual members beyond the limits of the similarity threshold set by clustering methods, or (III) incomplete assembly or annotation of genomes. It appears that many genes may be missed because of errors in the annotation process, even in well-studied species,[32,102] and this sort of "loss" should be examined carefully. In general, providing evidence for absence of a gene or family is a challenge for genome-wide studies of gene family evolution; however, in many cases the first two processes listed above do result in the complete loss of biological functions, even when there is evidence for the presence of ancient homologs. Previous analyses suggest that on average, *Drosophila* lose more genes *per* million years than do mammals.[2,25] A comparison of gene family content among the genomes of *Caenorhabditis elegans*, *Drosophila melanogaster*, humans, and pufferfish also supports a higher rate of gene family loss in invertebrates than vertebrates.[103] The difference in rates of loss may partially explain the large differences in gene number between these taxa.[103]

Future challenges

Evidence for the pervasiveness of evolution in gene copy number is difficult to obtain because it requires deeply sequenced whole genome coverage. Shallow sequencing coverage is inadequate because sequencing error, heterozygosity, and duplication cannot be distinguished. Consequently, recent duplicates are often collapsed or heterozygous single-copy genes are split into two apparent "paralogs." Inference of gene gain and loss can be especially problematic when the taxa under consideration suffer from heterogeneous sequence depth and/or annotation quality. For example, the near doubling of sequence coverage in the second release of the chimpanzee genome decreased the inferred number of duplications and losses between chimp and human by 8% (down from $\sim$14% in the initial chimp release to $\sim$6%).[25] In some cases using gene trees in conjunction with whole-genome alignments to determine syntenic regions where genes are expected to occur, may aid in distinguishing sequencing gaps from incorrect annotation or true losses (or gains).[32] Additionally, detailed cataloging of pseudogenes may provide evidence of absence; however, a recent analysis in *Drosophila* found that of 109 unambiguous gene losses in *D. melanogaster*, at most 18 had identifiable pseudogenes.[102]

Studies of gene family evolution would also benefit from additional theoretical work. Gene family evolution remains relatively understudied in comparison to analysis of orthologous sequences, in part because the theoretical expectations and mathematical machinery for orthologous sequence comparison are more mature.[31,104] Our recent efforts have put inferences about the role of selection in the evolution of gene family size in a more quantitative framework;[2,30,31] however, additional theoretical work would be beneficial. For instance, improvements including ways to deal with the large number of changes that accumulate over very long periods, and non-equilibrium models (*i.e.*, probability of birth $\neq$ death) would be very useful.

Finally, little experimental work has been done to characterize the rate of duplicative mutations and the

distribution of their fitness effects.[105] This question can best be addressed by complete genome sequencing of mutation accumulation lines. While large genome size combined with the large number of lines necessary to survey mutation makes this seem impractical in eukaryotes (except perhaps for yeast), the cost and speed of DNA sequencing is accelerating rapidly so that this experiment will be feasible for many model systems.

References

1. **Ohno, S.,** Evolution by gene duplication. Berlin, Springer-Verlag, 1970.
2. **Hahn, M. W., Demuth, J. P. and Han, S. G.,** Accelerated rate of gene gain and loss in primates. *Genetics* 2007. **177**: 1941–1949.
3. **Gu, X. and Li, W.-H.,** Higher rates of amino acid substitution in rodents than in humans. *Mol Phylogenet Evol* 1992. **1**: 211–214.
4. **Li, W.-H., Tanimura, M. and Sharp, P. M.,** An evaluation of the molecular clock hypothesis using mammalian DNA-sequences. *J Mol Evol* 1987. **25**: 330–342.
5. **Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T.,** *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 2005. **437**: 88–93.
6. **Britten, R. J.,** Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci USA* 2002. **99**: 13633–13635.
7. **Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H.,** *et al.* Global variation in copy number in the human genome. *Nature* 2006. **444**: 444–454.
8. **Dayhoff, M. O.,** Origin and evolution of protein superfamilies. *Fed Proc* 1976. **35**: 2132–2138.
9. **Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J.,** *et al.* The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 2007. **318**: 245–250.
10. **Muller, H. J.,** Bar duplication. *Science* 1936. **83**: 528–530.
11. **Chothia, C. and Lesk, A. M.,** The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986. **5**: 823–826.
12. **Hegyi, H. and Gerstein, M.,** The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 1999. **288**: 147–164.
13. **Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R.,** *et al.* Comparative genomics of the eukaryotes. *Science* 2000. **287**: 2204–2215.
14. **Thornton, J. W. and DeSalle, R.,** Gene family evolution and homology: genomics meets phylogenetics. *Annu Rev Genomics Hum Genet* 2000. **1**: 41–73.
15. **Doolittle, R. F.,** Similar amino acid sequences: chance or common ancestry? *Science* 1981. **214**: 149–159.
16. **Doolittle, R. F.,** The multiplicity of domains in proteins. *Annu Rev Biochem* 1995. **64**: 287–314.
17. **Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K.,** *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* 1997. **278**: 609–614.
18. **Long, M. Y. and Langley, C. H.,** Natural-selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* 1993. **260**: 91–95.
19. **Katju, V. and Lynch, M.,** On the formation of novel genes by duplication in the Caenorhabditis elegans genome. *Mol Biol Evol* 2006. **23**: 1056–1067.
20. Williams, D. M. and Forey, P. L. editors. Milestones in systematics. Boca Raton, CRC Press, 2004. p xvii, 290.
21. **Song, N., Morgan, J. M., Davis, G. B. and Durand, D.,** Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput Biol* 2008. **4**: e1000063.
22. **Tian, W. and Skolnick, J.,** How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 2003. **333**: 863–882.
23. **Hughes, A. L. and Friedman, R.,** Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. *J Mol Evol* 2004. **59**: 827–833.
24. **Kunin, V., Teichmann, S. A., Huynen, M. A. and Ouzounis, C. A.,** The properties of protein family space depend on experimental design. *Bioinformatics* 2005. **21**: 2618–2622.
25. **Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N. and Hahn, M. W.,** The evolution of mammalian gene families. *PLoS ONE* 2006. **1**: e85.
26. **Ouzounis, C. A., Coulson, R. M. R., Enright, A. J., Kunin, V. and Pereira-Leal, J. B.,** Classification schemes for protein structure and function. *Nat Rev Genet* 2003. **4**: 508–519.
27. **Krause, A.,** Large scale protein sequence clustering—not solved but solvable. *Curr Bioinform* 2006. **1**: 247–254.
28. **Zmasek, C. M. and Eddy, S. R.,** A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 2001. **17**: 821–828.
29. **Hahn, M. W.,** Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* 2007. **8**: R141.
30. **De Bie, T., Cristianini, N., Demuth, J. P. and Hahn, M. W.,** CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 2006. **22**: 1269–1271.
31. **Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. and Cristianini, N.,** Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* 2005. **15**: 1153–1160.
32. **Hahn, M. W., Han, M. V. and Han, S. G.,** Gene family evolution across 12 Drosophila genomes. *PLoS Genet* 2007. **3**: 2135–2146.
33. **Lynch, M. and Conery, J. S.,** The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 2003. **3**: 35–44.
34. **Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P. and Li, W.-H.,** Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol Biol Evol* 2002. **19**: 256–262.
35. **Lynch, M.,** The origins of genome architecture. Sunderland, MA, Sinaur Associates, Inc., 2007.
36. **Romero, D. and Palacios, R.,** Gene amplification and genome plasticity in prokaryotes. *Annu Rev Genet* 1997. **31**: 91–111.
37. **Mekalanos, J. J.,** Duplication and amplification of toxin genes in Vibrio cholerae. *Cell* 1983. **35**: 353–363.
38. **Kroll, J., Moxon, E. and Loynds, B.,** An ancestral mutation enhancing the fitness and increasing the virulence of Haemophilus influenzae type b. *J Infect Dis* 1993. **168**: 172–176.
39. **Cairns, J., Overbaugh, J. and Miller, S.,** The origin of mutants. *Nature* 1988. **335**: 142–145.
40. **Hendrickson, H., Slechta, E. S., Bergthorsson, U., Andersson, D. I. and Roth, J. R.,** Amplification-mutagenesis: evidence that ''directed'' adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proc Natl Acad Sci USA* 2002. **99**: 2164–2169.
41. **Slechta, E. S., Bunny, K. L., Kugelberg, E., Kofoid, E., Andersson, D. I.,** *et al.* Adaptive mutation: general mutagenesis is not a programmed response to stress but results from rare coamplification of dinB with lac. *Proc Natl Acad Sci USA* 2003. **100**: 12847–12852.
42. **Fogel, S., Welch, J. W., Cathala, G. and Karin, M.,** Gene amplification in yeast: CUP1 copy number regulates copper resistance. *Curr Genet* 1983. **7**: 347–355.
43. **Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O.,** *et al.* Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 2002. **99**: 16144–16149.
44. **Libuda, D. E. and Winston, F.,** Amplification of histone genes by circular chromosome formation in Saccharomyces cerevisiae. *Nature* 2006. **443**: 1003–1007.
45. **Vontas, J. G., Small, G. J. and Hemingway, J.,** Comparison of esterase gene amplification, gene expression and esterase activity in insecticide susceptible and resistant strains of the brown planthopper, Nilaparvata lugens (Stål). *Insect Mol Biol* 2000. **9**: 655–660.
46. **Field, L. M., Devonshire, A. L. and Forde, B. G.,** Molecular evidence that insecticide resistance in peach-potato aphids (Myzus persicae

Sulz.) results from amplification of an esterase gene. *Biochem J* 1988. **251**: 309–312.
47. **Mouches, C., Pasteur, N., Berge, J. B., Hyrien, O. Raymond, M.,** *et al.* Amplification of an esterase gene is responsible for insecticide resistance in a California Culex mosquito. *Science* 1986. **233**: 778–780.
48. **Alt, F. W., Kellems, R. E., Bertino, J. R. and Schimke, R. T.,** Selective multiplication of dihydrofolate reductase genes in methotrexate-resistant variants of cultured murine cells. *J Biol Chem* 1978. **253**: 1357–1370.
49. **Albertson, D. G.,** Gene amplification in cancer. *Trends Genet* 2006. **22**: 447–455.
50. **Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A.,** *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987. **235**: 177–182.
51. **Francino, M. P.,** An adaptive radiation model for the origin of new gene functions. *Nat Genet* 2005. **37**: 573–577.
52. **Zhang, J. Z.,** Evolution by gene duplication: an update. *Trends Ecol Evol* 2003. **18**: 292–298.
53. **Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I. and Koonin, E. V.,** Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res* 2001. **11**: 555–565.
54. **Gevers, D., Vandepoele, K., Simillion, C. and Van de Peer, Y.,** Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* 2004. **12**: 148–154.
55. **Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, D.,** *et al.* Dynamic evolution of the innate immune system in Drosophila. *Nat Genet* 2007. **39**: 1461–1468.
56. **Hughes, A. L.,** The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B* 1994. **256**: 119–124.
57. **Tanaka, T. and Nei, M.,** Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Mol Biol Evol* 1989. **6**: 447–459.
58. **Rispe, C., Kutsukake, M., Doublet, V., Hudaverdian, S., Legeai, F.,** *et al.* Large gene family expansion and variable selective pressures for Cathepsin B in Aphids. *Mol Biol Evol* 2008. **25**: 5–17.
59. Rhesus Macaque Genome Sequencing and Analysis Consortium. Evolutionary and biomedical insights from the Rhesus Macaque genome. *Science* 2007. **316**: 222–234.
60. **Birchler, J. A. and Veitia, R. A.,** The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* 2007. **19**: 395–402.
61. **Seoighe, C. and Gehring, C.,** Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. *Trends Genet* 2004. **20**: 461–464.
62. **Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M.,** *et al.* Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* 2005. **102**: 5454–5459.
63. **Grus, W. E., Shi, P., Zhang, Y-p. and Zhang, J.,** Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. *Proc Natl Acad Sci USA* 2005. **102**: 5767–5772.
64. **Go, Y., Satta, Y., Takenaka, O. and Takahata, N.,** Lineage-specific loss of function of bitter taste receptor genes in humans and nonhuman primates. *Genetics* 2005. **170**: 313–326.
65. **McBride, C. S., Arguello, J. R. and O'Meara, B. C.,** Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* 2007. **177**: 1395–1416.
66. **Niimura, Y. and Nei, M.,** Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE* 2007. **8**: e708.
67. **Nei, M., Zhang, J. and Yokoyama, S.,** Color vision of ancestral organisms of higher primates. *Mol Biol Evol* 1997. **14**: 611–618.
68. **Gilad, Y., Wiebe, V., Przeworski, M., Lancet, D. and Pääbo, S.,** Correction: loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol* 2007. **5**: e148.
69. **Gilad, Y., Bustamante, C. D., Lancet, D. and Pääbo, S.,** Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am J Hum Genet* 2003. **73**: 489–501.
70. **Olson, M. V.,** When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* 1999. **64**: 18–23.
71. **Stedman, H. H., Kozyak, B. W., Nelson, A., Thesier, D. M., Su, L. T.,** *et al.* Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 2004. **428**: 415–418.
72. **Ajit, V.,** Loss of N-glycolylneuraminic acid in humans: mechanisms, consequences, and implications for hominid evolution. *Am J Phys Anthropol* 2001. **116**: 54–69.
73. **Wang, X., Grus, W. E. and Zhang, J.,** Gene losses during human origins. *PLoS Biol* 2006. **4**: e52.
74. **Karro, J. E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N.,** *et al.* Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucl Acids Res* 2007. **35**: D55–D60.
75. **Petrov, D. A., Chao, Y. C., Stephenson, E. C. and Hartl, D. L.,** Pseudogene evolution in Drosophila suggests a high rate of DNA loss. *Mol Biol Evol* 1998. **15**: 1562–1567.
76. **Moran, N. A.,** Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 2002. **108**: 583–586.
77. **Pal, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G.,** *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* 2006. **440**: 667–670.
78. **Marais, G., Calteau, A. and Tenaillon, O.,** Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica* 2008. **134**: 205–210.
79. **Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L.,** *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 2005. **309**: 1242–1245.
80. **Niimura, Y. and Nei, M.,** Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene* 2005. **346**: 23–28.
81. **McBride, C. S.,** Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia. *Proc Natl Acad Sci USA* 2007. **104**: 4996–5001.
82. **Amiri, H., Davids, W. and Andersson, S. G. E.,** Birth and death of orphan genes in Rickettsia. *Mol Biol Evol* 2003. **20**: 1575–1587.
83. **Domazet-Loso, T. and Tautz, D.,** An evolutionary analysis of orphan genes in Drosophila. *Genome Res* 2003. **13**: 2213–2219.
84. **Schmid, K. J. and Aquadro, C. F.,** The evolutionary analysis of ''orphans'' from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* 2001. **159**: 589–598.
85. **Hall, C., Brachat, S. and Dietrich, F. S.,** Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. *Eukaryot Cell* 2005. **4**: 1102–1115.
86. **Mower, J. P., Stefanovic, S., Young, G. J. and Palmer, J. D.,** Plant genetics: gene transfer from parasitic to host plants. *Nature* 2004. **432**: 165–166.
87. **Dunning Hotopp, J. C., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P.,** *et al.* Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 2007. **317**: 1753–1756.
88. **Begun, D. J., Lindfors, H. A., Kern, A. D. and Jones, C. D.,** Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba Drosophila erecta clade. *Genetics* 2007. **176**: 1131–1137.
89. **Wilson, G. A., Bertrand, N., Patel, Y., Hughes, J. B., Feil, E. J.,** *et al.* Orphans as taxonomically restricted and ecologically important genes. *Microbiology* 2005. **151**: 2499–2501.
90. **Bosch, N., Caceres, M., Cardone, M. F., Carreras, A., Ballana, E.,** *et al.* Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. *Hum Mol Genet* 2007. **16**: 2572–2582.
91. **Johnson, M. E., Viggiano, L., Bailey, J. A., Abdul-Rauf, M., Goodwin, G.,** *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* 2001. **413**: 514–519.
92. **Popesco, M. C., MacLaren, E. J., Hopkins, J., Dumas, L., Cox, M.,** *et al.* Human lineage-specific amplification, selection, and neuronal expression of DUF1220 Domains. *Science* 2006. **313**: 1304–1307.
93. **Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. and van Roy, F.,** A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol* 2005. **22**: 2265–2274.
94. **Robinson-Rechavi, M., Maina, C. V., Gissendanner, C. R., Laudet, V. and Sluder, A.,** Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J Mol Evol* 2005. **60**: 577–586.

95. **Ochman, H., Lawrence, J. G. and Groisman, E. A.,** Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000. **405**: 299–304.
96. **Forterre, P., de la Tour, C. B., Philippe, H. and Duguet, M.,** Reverse gyrase from hyperthermophiles—probable transfer of a thermoadaptation trait from Archaea to Bacteria. *Trends Genet* 2000. **16**: 152–154.
97. **Andersson, J. O.,** Lateral gene transfer in eukaryotes. *Cell Mol Life Sci* 2005. **62**: 1182–1197.
98. **Furner, I. J., Huffman, G. A., Amasino, R. M., Garfinkel, D. J., Gordon, M. P.,** *et al*. An Agrobacterium transformation in the evolution of the genus Nicotiana. *Nature* 1986. **319**: 422–427.
99. **Richardson, A. O. and Palmer, J. D.,** Horizontal gene transfer in plants. *J Exp Bot* 2007. **58**: 1–9.
100. **Hittinger, C. T., Rokas, A. and Carroll, S. B.,** Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci USA* 2004. **101**: 14144–14149.
101. **Payne, S. H. and Loomis, W. F.,** Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Eukaryot Cell* 2006. **5**: 272–276.
102. **Costello, J. C., Han, M. V. and Hahn, M. W.,** Limitations of pseudogenes in identifying gene losses. *Lect Notes Bioinform* 2008. **5267**: 14–25.
103. **Hughes, A. L. and Friedman, R.,** Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proc R Soc B Biol Sci* 2004. **271**: S107–S109.
104. **Li, W.-H.,** Molecular evolution. Sunderland, MA, Sinaur Associates, Inc., 1997.
105. **Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R.,** *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 2008. **105**: 9272–9277.