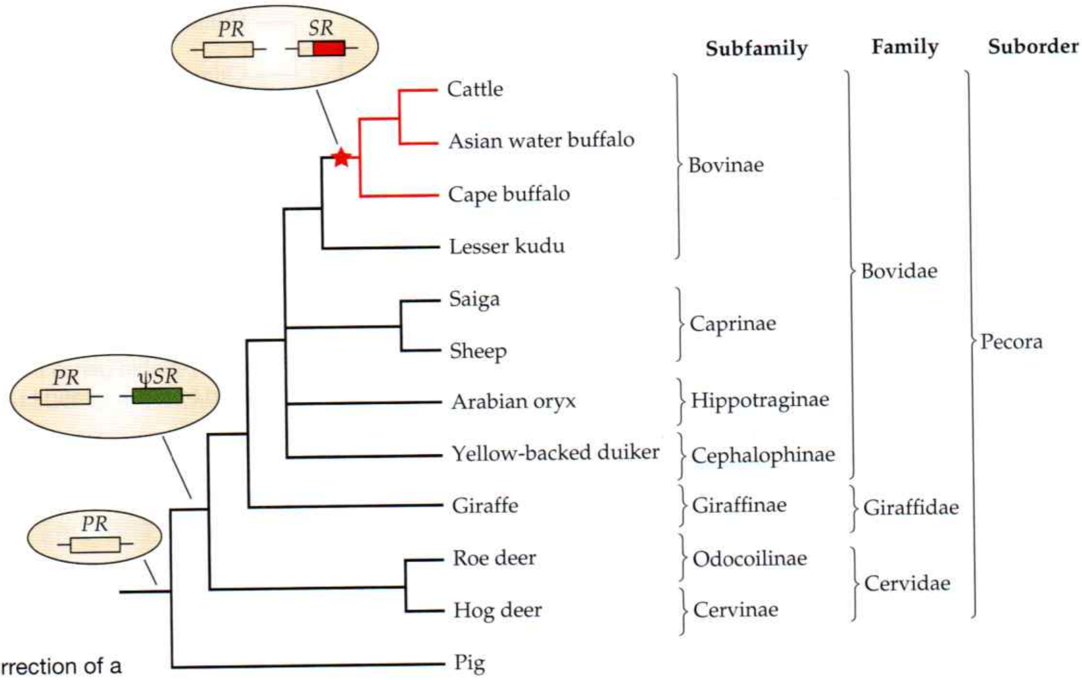


Figure 7.26 The resurrection of a ribonuclease pseudogene by gene conversion in a bovine lineage. Species in which the seminal ribonuclease protein is functional are marked with a plus sign (+). In the ancestor of the pecorans, the pancreatic ribonuclease (*PR*) gene (empty box) was triplicated to yield a nonfunctional seminal ribonuclease (ψ*SR*) pseudogene (green box) and a functional cerebral ribonuclease gene (not shown). In the ancestor of cattle and buffaloes, the 5′ end of the pseudogene was converted by the functional gene and became a functional gene (red star) encoding seminal ribonuclease (*SR*). Species: cattle, *Bos taurus*; Asian water buffalo, *Bubalus bubalis*; Cape buffalo, *Syncerus caffer*; lesser kudu, *Tragelaphus imberbis*; saiga, *Saiga tatarica*; sheep, *Ovis aries*; Arabian oryx, *Oryx leucoryx*; yellow-backed duiker, *Cephalophus sylvicultor*; giraffe, *Giraffa camelopardalis*; roe deer, *Capreolus capreolus*; hog deer, *Axis porcinus*; pig, *Sus scrofa*. (Modified from Graur and Li 2000.)

et al. 1996). The gene conversion event involved only a small region at the 5′ end of the gene, including about 70 nucleotides in the untranslated region. Interestingly, this conversion not only removed a deletion that caused a frameshift in the reading frame, but also restored two amino acid residues that are vital for the proper functioning of ribonuclease: histidine at position 12 and cysteine at position 31.

Thanks to gene conversion, death (as far as genes are concerned) does not by necessity connote finality. Thus, while pseudogenes are nonfunctional by definition, genomic extinction is not their only fate. They may, albeit very rarely, take part in the evolution of functional genes through gene conversion, unequal recombination, and transposition (Chapter 9).

We note that gene conversion is a mutational process, i.e., the proximity of a gene to a pseudogene may not only spell rebirth for the pseudogene, but can also spell death for the gene. One such example of gene death by gene conversion concerns *CYP21A2* (cytochrome P450 gene, family 21, subfamily A, polypeptide 2). In humans, *CYP21A2* is a 10-exon gene that is located on chromosome 6 in a region in which many major histocompatibility and complement genes are interspersed with one another. The gene has a paralogous nonprocessed pseudogene in the vicinity. Interestingly, many mammals have one functional copy and one nonfunctional copy of the *CYP21A2* gene; however, the nonfunctionalization event giving rise to the pseudogene occurred independently in many of the various lineages. Thus, for instance, the ortholog of the human functional gene is a pseudogene in mouse, and the ortholog of the human pseudogene is a functional gene in mouse.

Hundreds of mutations in *CYP21A2* have been characterized in the clinical literature (mostly causing a disease called congenital adrenal hyperplasia), and about 75% of them have been found to be due to gene conversion (Mornet et al. 1991). In at least one case, by using a de novo mutation in an individual, it was possible to pinpoint an intrachromatid gene conversion event involving 390 nucleotides in the maternal chromosome as the cause of nonfunctionalization (Collier et al. 1993).

The relative roles of gene conversion and unequal crossing over

As a mechanism for concerted evolution, gene conversion appears to have several advantages over unequal crossover. First, unequal crossover generates changes in the number of repeated genes within a family, which may sometimes cause a significant dosage imbalance. For example, the deletion of one of the two α-globin genes following an unequal crossover gives rise to a mild form of α-thalassemia in homozygotes (Lupski 1998). Gene conversion, on the other hand, causes no change in gene number.

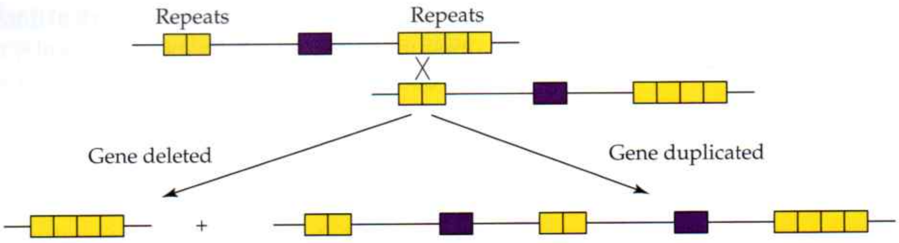


Figure 7.27 Crossover involving dispersed repeats (yellow boxes). The purple box denotes a unique gene. In the crossover event, the unique gene is deleted in one chromosome and duplicated in the other. (Modified from Graur and Li 2000.)

Second, gene conversion can act as a correction mechanism not only on tandem repeats but also on dispersed repeats within a chromosome (Jackson and Fink 1981; Klein and Petes 1981), between homologous chromosomes (Fogel et al. 1978), or between nonhomologous chromosomes (Scherer and Davis 1980; Ernst et al. 1982). In contrast, unequal crossover is severely restricted when repeats dispersed on nonhomologous chromosomes are involved. It can probably act effectively on nonhomologous chromosomes only if the repeated genes are located on the telomeric parts of the chromosome (the ends of the chromosome arms), as in the case of rRNA genes in humans and apes, but will be greatly restricted if the dispersed repeats are located in the middles of chromosomes, as in the case of rRNA genes in mice, lizards, and *Drosophila melanogaster.* If the repeats are dispersed on a chromosome, unequal crossover can result in the deletion or duplication of the genes that are located between the repeats. For example, **Figure 7.27** shows a hypothetical case of unequal crossover between two repeated clusters, resulting in the deletion of a unique gene in one chromosome and a corresponding duplication in the other. Either one or both chromosomes could have a deleterious effect on their carriers.

Third, gene conversion can be biased, i.e., have a preferred direction. Experimental data from fungi have shown that bias in the direction of gene conversion is common and often strong (Lamb and Helmi 1982), and theoretical studies have shown that even a small bias can have a large effect on the probability of fixation of repeated mutants (Nagylaki and Petes 1982; Walsh 1985).

For the above reasons, some authors (Baltimore 1981; Dover 1982; Nagylaki and Petes 1982) have proposed that gene conversion plays a more important role in concerted evolution than unequal crossover. This is probably true for dispersed repeats, because in this case gene conversion can act more effectively than unequal crossover. It is also probably true for small multigene families (e.g., the duplicated α-globin genes in humans), because in such families unequal crossover may cause severe adverse effects. In large families of tandemly repeated sequences, however, unequal crossover may be as acceptable a process as gene conversion. Indeed, in such cases, unequal crossover may be faster and more efficient than gene conversion in bringing about concerted evolution, for several reasons.

First, in such families, the number of repeats apparently can fluctuate greatly without causing significant adverse effects. This is suggested by the observations that the number of RNA-specifying genes in *Drosophila* varies widely among individuals of the same species and among species (Ritossa et al. 1966; Brown and Sugimoto 1973). Moreover, in humans, several families of tandem repeats that exhibit extraordinary degrees of variation in copy number have been found (Nakamura et al. 1987). Second, in a gene conversion event, usually only a small region (the heteroduplex region) is involved, whereas in unequal crossover the number of repeats that are exchanged between the chromosomes can be very large. For example, in yeast, a single unequal crossover event was shown to involve on average seven repeats of rRNA genes, i.e., ~20,000 bp (Szostak and Wu 1980), whereas a gene conversion track may not exceed 1,500 bp (Curtis and Bender 1991). Obviously, the larger the number of repeats exchanged, the higher the rate of concerted evolution will be (Ohta 1983). In some cases, this advantage of unequal crossover may be large enough to offset those of gene conversion. Finally, the empirical data show that in some organisms (e.g., yeast), unequal crossover occurs more frequently than nonallelic gene conversion. Of course, the observed lower rate of gene conversion might have been due to a detection bias, for it is generally much easier to detect unequal crossover than gene conversion.

Factors Affecting Concerted Evolution

How cohesively the members of a repeated-sequence family evolve together depends on several factors, including the number of repeats (i.e., the size of the gene family), the arrangement of the repeats, the structure of the repeated unit, the functional constraints imposed on the repeated unit, the mechanisms of concerted evolution, and the selective and nonselective processes at the population level.

Number of repeats

It is quite easy to see that the rate of concerted evolution is dependent on the number of repeats. For example, if there are only two repeats on a chromosome, a single intrachromosomal gene conversion will lead to homogeneity of the repeats on the chromosome. On the other hand, when there are more than two repeats on the chromosome, more than a single conversion may be required to homogenize the sequences.

Smith (1974, 1976) seems to be the first author to have conducted a quantitative study of the effect of family size on the rate of homogenization in a multigene family. His simulation study indicated that the number of unequal crossover events required for the fixation of a variant repeat in a single chromosomal lineage increases roughly with n^2, where n is the number of repeats on the chromosome.

Arrangement of repeats

There are, roughly speaking, two types of arrangement of repeated units. In some gene families, the members are highly dispersed all over the genome. One example is the human *Alu* family, whose approximately one million members are interspersed with single-copy sequences throughout the genome (Chapter 11). This type of arrangement is the least favorable for concerted evolution, because it greatly reduces the chance of unequal crossover and gene conversion and because unequal crossover will often lead to disastrous genetic consequences. Thus, the high similarities among *Alu* sequences are most probably due to relatively recent amplification events of source sequences (Chapter 9) rather than to concerted evolution.

In the second type of arrangement, all members of a family are clustered either in a single tandem array or in a small number of tandem arrays located on different chromosomes. This arrangement is the most favorable for unequal crossover and gene conversion. If the repeats are located on more than one chromosome, the rate of unequal crossover is greatly reduced, unless the clusters occur at the ends of chromosome arms (as in the case of the rDNA family in humans). Moreover, the rate of gene conversion would also be reduced. However, Ohta and Dover (1983) have shown that such a reduction in gene conversion rate has only a minor effect on the extent of identity between genes, unless the conversion rate between genes on nonhomologous chromosomes becomes very low, or unless the number of nonhomologous chromosomes on which gene family members reside is large.

Structure of the repeat unit

The structure of the repeat unit refers to the numbers and sizes of coding regions (i.e., exons) and noncoding regions (i.e., introns and spacers) within the repeat unit. As noncoding regions generally evolve rapidly, it is difficult to maintain a high degree of similarity among the repeats if each repeat contains large or numerous noncoding regions. We note that homogeneity and concerted evolution go hand in hand, since both unequal crossover and gene conversion depend on sequence similarity for misalignment of repeats. Thus, the higher the homogeneity among the repeats in a family, the higher the rates of unequal crossover and gene conversion.

Zimmer et al. (1980) estimated that in the great apes, the rate of concerted evolution in the α-globin gene region is 50 times higher than that in the β-globin gene region. They suggested that the rate in the β region has been greatly reduced because the introns and flanking sequences are highly divergent between the two β genes. It is interesting to note that the β genes have introns that are several times longer than those of the α genes, and that the intergenic region between the two β genes is 2,400 bases longer than that between the two α genes. Indeed, Zimmer et al. (1980) suggested that the larger introns and intergenic region in the β genes arose as a response to selection against unequal crossover, which may produce a single homogenized gene out of the β- and δ-globin genes (called hemoglobin Lepore), whose expression is under the control of the δ promoter and is deleterious in the homozygous state.

We note, however, that qualitative (as opposed to quantitative) arguments concerning putative advantages associated with protection against mutational events (e.g., avoidance of pretermination codons, prevention of crossover events) are usually highly exaggerated, because the selective advantage for a reduction in the rate of a mutational event would at most be as large as the mutation rate itself. Assuming that mutational events occur at rates of 10^{-5} to 10^{-9}, the selective advantage would be insignificant except in taxa with enormous effective population sizes. Thus, the large introns and intergenic regions might have arisen by chance rather than by selection. It is possible that the introns and the intergenic regions were already large before the divergence of the apes, and that this promoted the divergence between the two β genes rather than vice versa.

Functional requirements and selection

Let us first consider two extreme situations. One is that the function has an extremely stringent structural requirement, often requiring large amounts of the same gene product (dose repetitions). The rRNA genes and the histone genes are well-known examples. The other extreme is that the function requires a large amount of diversity. The immunoglobulin and histocompatibility genes belong to this category.

In general, the rate of concerted evolution is expected to be higher in the former type than in the latter type of families. In rRNA genes, purifying selection will tend to eliminate new variants and promote homogeneity, which in turn will facilitate unequal crossover and gene conversion among members of multigene families, thus accelerating the process of concerted evolution. In immunoglobulin genes, on the other hand, an individual with many identical copies owing to concerted evolution would be at a severe disadvantage, as its arsenal of immunoglobulins against pathological antigens would be limited. Thus, concerted evolution may not play an important role in the evolution of such families.

A general feature of concerted evolution is that it cannot persist for long periods of time in the absence of purifying selection. This is because both gene conversion and unequal crossing over require a high degree of sequence similarity between the duplicate genes so that pairing between two paralogous sequences can occur. Without selection, sequence similarity tends to decrease with time because of the accumulation of mutations. Teshima and Innan (2004) studied concerted evolution under a mathematical model that allows mutation and gene conversion, but no selection. They found that concerted evolution of duplicate genes can persist for only short periods of time. As seen in Figure 7.28, the average sequence difference (d) between two duplicated sequences is 0 at the time of the gene duplication event, but it will increase with time to reach the equilibrium value (d_0) that is determined by the opposing effects of mutation and gene conversion (phase I). Then, d will fluctuate around the equilibrium value because of stochastic effects (phase II) until chance carries it over a threshold value (d_t), above which gene conversion can no longer occur. From this point onward, d will increase monotonically with time (phase III). In this model, a clear-cut threshold value is assumed, such that the probability of gene conversion suddenly becomes 0 when d exceeds the threshold. In reality the probability is likely to decrease gradually as d increases. Under this situation, the transition from phase II to phase III is less

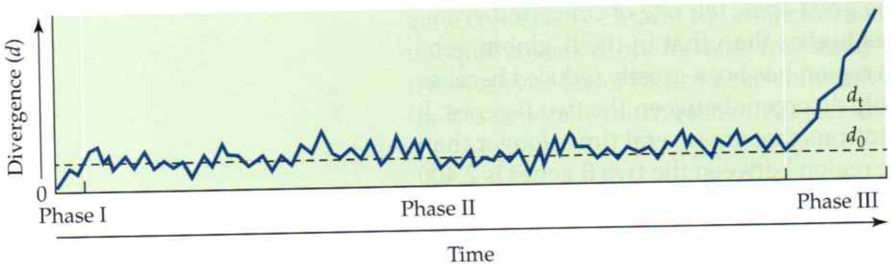


Figure 7.28 A schematic example of the changes in average sequence divergence (d) between two paralogs after a gene duplication event at time 0. In phase I, d tends to increase until it reaches an equilibrium value (d_0; dashed line). In phase II, d fluctuates around the equilibrium value. In phase III, d happens to exceed a threshold (d_t), above which neither gene conversion can occur, so d increases monotonically with time. The time duration of phase II follows an exponential distribution that depends on the rate of gene conversion, the length of the conversion tract, and the rate of mutation. (Modified from Teshima and Innan 2004.)

sharp but will eventually occur. Thus, without purifying selection to maintain sequence similarity, concerted evolution between duplicated genes cannot continue for a very long time.

For protein-coding genes, functional constraints at the protein level will retard the rate of divergence between sequences, so concerted evolution can persist for a longer time than for sequences without function. This is particularly true for genes coding for highly constrained proteins such as histones. However, even for protein-coding genes, sequence divergence can occur at degenerate sites where synonymous substitution can occur. In this case, the strength of codon usage bias can make a difference in the persistence of concerted evolution, with genes that exhibit highly biased codon usage being more disposed to concerted evolution than genes that exhibit no codon preference (Gao and Innan 2004).

Lin et al. (2006) pointed out that purifying selection is a constant force that operates in each and every generation, whereas gene conversion is unlikely to occur with any constancy or regularity. Moreover, purifying selection can occur without gene conversion, whereas concerted evolution conversion cannot continue for long without purifying selection. Thus, purifying selection is more potent in maintaining sequence homogeneity between duplicated genes than either gene conversion or unequal crossover.

Finally, we note that concerted evolution by unequal crossing over may be affected by centripetal selection, i.e., purifying selection against too many or too few repeats. Specifically, unequal crossing over is known to create a large variation in the number of repeats among individuals in a population. If a certain number of copies is required, i.e., if an optimum copy-number range exists, then centripetal selection may become an important force shaping concerted evolution patterns.

Population size

Population size affects the rate of concerted evolution because concerted evolution requires not only the horizontal spread of genetic variation among members of a gene family, but also the fixation of such homogeneous variants within the population. Obviously, the time required for a variant to be eliminated from a population or to become fixed in a population is dependent on the population size (Chapter 2).

Positive natural selection will accelerate the process of concerted evolution because the rate and probability of fixation for a variant favored by natural selection will be larger than those for selectively neutral variants. The effect of biased gene conversion on the evolution of multigene families would be similar to that of positive selection, albeit somewhat weaker. In addition, biased gene conversion will be more effective when the number of repeats is large (Walsh 1985). Both natural selection and biased gene conversion work more effectively in large populations than in small ones, because the effect of random genetic drift decreases with population size.

Evolutionary Implications of Concerted Evolution

Concerted evolution allows the spread of a variant repeat to all gene family members. This capability has profound evolutionary consequences. In this section we discuss the effects of concerted evolution on the spread of advantageous mutations, the rate of divergence between duplicate genes, and the generation of genic variation.

Spread of advantageous mutations

Through concerted evolution, an advantageous mutant can spread rapidly and replace all other repeats within a gene family. We note that the selective advantage that a single variant can confer on an organism is usually very small. The advantage would, however, be greatly amplified if the mutation were to spread within the genome. Thus, through concerted evolution, a small selective advantage can become a great advantage. In this respect, concerted evolution surpasses independent evolution of individual gene family members (Arnheim 1983; Walsh 1985).

Arnheim (1983) compared the evolution of RNA polymerase I transcriptional control signals with that of RNA polymerase II transcriptional control signals. RNA polymerase I transcribes rRNA genes, whereas RNA polymerase II transcribes protein-coding genes (Chapter 1). RNA polymerase I transcriptional control signals appear to have evolved much faster than the signals for RNA polymerase II. For example, in cell-free transcription systems, a mouse rDNA clone does not work in a human cell extract, but clones of protein-coding genes from astonishingly diverse species can be transcribed in heterologous systems (e.g., silkworm genes in human cell extracts, and mammalian genes in yeast). Arnheim (1983) argues that in the case of transcription units for RNA polymerase I, mutations that favorably affect transcription initiation were propagated throughout the rDNA multigene family as a consequence of concerted evolution, and they could become species-specific. On the other hand, in the case of transcription units for RNA polymerase II, advantageous mutations affecting transcription initiation that occur in any one gene would not be expected to be propagated throughout all genes, for they belong to many different families.

Retardation of paralogous gene divergence

The traditional view concerning the creation of a new function is that a gene duplication event occurs, and one of the two resultant genes gradually diverges and becomes a new gene. It is now clear that the process may not be as simple as previously assumed. As long as the degree of divergence between the two genes is not large (as is the case immediately after the duplication event), one copy may be deleted by unequal crossover or converted by the sequence of the second copy by gene conversion. In the former case, an additional duplication would be required to create a new redundant copy, while in the latter case divergence must start again from scratch. Thus, divergence of duplicate genes may proceed much more slowly than traditionally thought.

Generation of genic variation

From an evolutionary point of view, there is an analogy between the evolution of multigene families and the evolution of subdivided populations. We may regard each repeat in a multigene family as a deme in a subdivided population. The transfer of information between repeats is then equivalent to the migration of genes or individuals between demes. It is well known that migration reduces the amount of genetic difference between demes but increases the amount of genic variation (i.e., the number of alleles) in a deme. Similarly, transfer of information between repeats will reduce the genetic difference between repeats but will increase the amount of genic variation at a locus (Ohta 1983, 1984; Nagylaki 1984). Indeed, some loci in the mouse major histocompatibility complex are highly polymorphic, with as many as 50 alleles being observed at a locus, and it has been suggested that the high polymorphism is due to concerted evolution (Weiss et al. 1983). An alternative explanation is that the alleles have persisted in the population for very long periods of time (Figueroa et al. 1988), probably being maintained by overdominant selection (Hughes and Nei 1989). Of course, these two mechanisms are not mutually exclusive, and they may both operate at these loci.

Methodological pitfalls due to concerted evolution

It has been customary to assume that, following a gene duplication, the two resultant genes diverge monotonically with time. Under this assumption, as we have previously shown, it is rather simple to infer the time of the duplication event (p. 275). An unfortunate feature of concerted evolution is that it erases the record of molecular divergence during the evolution of paralogous sequences. Thus, when dealing with very similar paralogous sequences from a species, it is usually impossible to distinguish between two possible alternatives: (1) the sequences have only recently diverged from one another by duplication, or (2) the sequences have evolved in concert. One way to distinguish between the alternatives is to use a phylogenetic approach. For example, the two α-globin genes in humans are almost identical to one another. Initially they were thought to have duplicated quite recently and that there had not been sufficient time for them to diverge in sequence. However, duplicated α-globin genes were also discovered in distantly related species, and so one had to assume either that multiple gene duplication events occurred independently in a great number of evolutionary lineages, or that the two genes are quite ancient, having been duplicated once in the common ancestor of these organisms, their antiquity subsequently obscured by concerted evolution.

Under concerted evolution, gene duplications appear younger than they really are. Phylogenetic reconstructions based on sequence comparisons can only go back to the last erasure of the evolutionary history. We must therefore use taxonomic information concerning the distribution of duplicated genes versus unduplicated ones to infer the time of gene duplication. In large multigene families, gene correction events are expected to occur frequently, and in such cases it will be even more difficult to trace the evolutionary relationships among the family members. Thus, concerted evolution should be taken into account when attempting to reconstruct the evolutionary history of paralogous genes. Failure to consider this possibility may result in faulty phylogenetic reconstructions.

Positive selection or biased gene conversion? The curious histories of HAR1 and FXY

A particularly intriguing methodological pitfall concerns the confusion between GC-biased gene conversion and selection (see Galtier and Duret 2007; Galtier et al. 2009). GC-biased gene conversion (Marais 2003) is a recombination-associated segregation distortion favoring GC-rich sequences over AT-rich sequences. A trivially expected effect of GC-biased gene conversion is an increase in the GC content of those DNA sequences undergoing conversion. This feature of GC-biased gene conversion has the unfortunate result of causing scientists to identify genomic regions under positive selection where none exist. Moreover, it serves to illustrate a particularly "unintelligent" characteristic of the evolutionary process, i.e., that a mutational process (gene conversion) not only does not promote adaptation but may also lead to the fixation of deleterious traits.

Pollard et al. (2006) proposed an elegant approach for detecting lineage-specific instances of positive selection. They were particularly interested in human sequences that experienced positive selection as candidates for adaptations involved in what makes us human. They first sought orthologous regions that exhibit high levels of conservation in all vertebrates but were very different in humans. Among these regions, they discovered 49 regions in which the substitution rate was significantly elevated in humans in comparison to the rates in other organisms. These regions were called "human accelerated regions" (HARs) and were christened *HAR1–HAR49*. The 118 bp *HAR1*, for instance, showed only two differences between chimpanzees and chickens (>300 million years of divergence) but 18 between chimpanzees and humans (~6 million years of divergence). The strong conservation of HARs among nonhuman vertebrates clearly indicates that they are under functional constraints. The sudden acceleration in substitution rates in the human lineage cannot be explained simply by loss of function and relaxation of selection, because the amount of change in the human lineage is much higher than expected for a neutrally evolving sequence—the average human-chimpanzee divergence in noncoding sequences is less than 2%. HARs are, therefore, good candidates for having evolved under positive selection in humans.

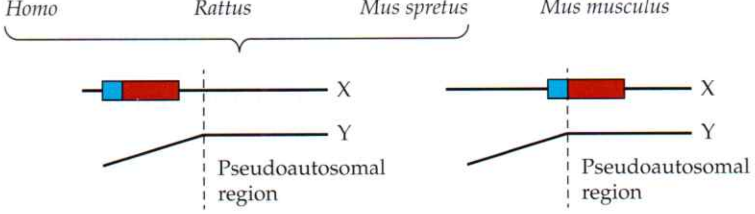
We note, however, that forces other than natural selection can lead to an increased nucleotide substitution rate. First, the rate of mutation can vary along genomes; however, the intragenomic variation in mutation rates that has been reported so far is relatively small (about twofold) compared with the 20-fold increase in the substitution rate in *HAR1*. The second possible explanation is biased gene conversion, in particular GC-biased gene conversion. From an evolutionary point of view, GC-biased gene conversion will yield results that are almost indistinguishable from positive selection (Galtier 2004). A GC-biased gene conversion event will lead to the transmission and eventual fixation of AT → GC mutations. The result will mimic the consequences of an episode of adaptive evolution, including an increased substitution rate.

How can we distinguish between positive selection and GC-biased gene conversion? This question is particularly pertinent to the study of regions that do not code for proteins, for in protein-coding regions we may examine the ratios of nonsynonymous to synonymous substitution as indicators of selection regimes. First, we note that although under GC-biased gene conversion, GC-rich sequences are favored, under positive selection there is no a priori reason why advantageous alleles should systematically be GC-rich. Notably, all 18 nucleotide substitutions observed in *HAR1* are AT → GC changes. Second, *HAR1* is located in a highly recombining region. This is a necessary condition for GC-biased gene conversion, but not for any known type of selection. Third, a mutational process such as GC-biased gene conversion has no regard for function. Indeed, the region of rapid evolution in *HAR1* is not restricted to the 118 bp region, which is conserved in all vertebrates with the exception of humans, but extends to flanking sequences.

A spectacular "experiment of nature" illustrating the effects of biased-gene conversion can be seen in the case of the evolution of the *FXY* gene in mammals (**Figure 7.29**). This gene is X-linked in human, rat, and short-tailed mouse (*Mus spretus*) but was recently translocated to another location on the X chromosome in the house mouse, *M. musculus*. In the house mouse, it now overlaps the boundary between the X-specific region and the pseudoautosomal region, with the 5′ end of the gene (exons 1–3) being located in the X-specific region, which does not recombine in males, and the 3′ end (exons 4–10) in the pseudoautosomal region, a short segment of homology between the X and Y chromosomes, which is a highly recombining region in both males and females. As a consequence of the move, the 3′ *FXY* sequence experienced a sudden increase in recombination rate, followed by a dramatic increase in GC content and in inferred substitution rates. No such acceleration was observed in the 5′ region.

At the protein level, the consequences of the translocation are equally dramatic. The 667-amino-acid protein is highly conserved between mammals (excluding *M. musculus*). The human and *M. spretus* sequences, which diverged more than 80 million years ago, differ by just 6 amino acid replacements. The rat and *M. spretus* sequences, which diverged 10–12 million years ago, differ by 5 amino acid replacements. The house mouse sequence, however, has accumulated as many as 28 amino acid replacements since the divergence from *M. spretus*, roughly 1–3 million years ago (Kurzweil et al. 2009). This corresponds to more than a 100-fold increase in the rate of amino acid replacements in the *M. musculus* lineage. Interestingly, all of these amino acid replacements have occurred in the regions encoded by the 3′ exons and were caused by AT → GC substitutions. Obviously, this elevated substitution rate has nothing to do with selection or adaptation. If directional selection had been acting on the protein sequence, then silent sites (most third-codon positions and introns) should have remained unaffected, but they did not. The estimated numbers of synonymous substitutions on the

Figure 7.29 Evolutionary history of *FXY* genes in four mammalian species. In most mammals the *FXY* gene is located in the X-specific region, which cannot recombine in males (diagram below tree). In the house mouse, *Mus musculus*, the gene was translocated to a new position on the X chromosome. This position overlaps the pseudoautosomal boundary (vertical dashed line), with exons 1–3 (blue) located in the X-specific region and exons 4–10 (red) located in the pseudoautosomal region. For each branch, the numbers of nonsynonymous substitutions that have occurred in the 5′ and 3′ ends of the gene are given on the left and right sides of a forward slash, respectively. A huge increase in the rate of amino acid replacement can be seen to have occurred in the 3′ end of the *FXY* gene in the *M. musculus* lineage. (Data from Galtier and Duret 2007.)



M. musculus branch are 1 for the 5′ end of the gene and 163 for the 3′ end. Moreover, if selective pressures were at work, they should have affected the entire gene, and not only its 3′ end. Finally, there is no reason to assume that all adaptive changes will be caused by AT → GC substitutions.

An important lesson from *HAR1* and *FXY* is that evolution is neither intelligent nor monotonically directional with respect to adaptation. GC-biased gene conversion can theoretically overcome purifying selection and lead to the fixation of deleterious AT → GC mutations. Conversely, it can overcome positive selection and erase an advantageous GC → AT mutation. Despite initial claims, it is unlikely that *HAR1* and *FXY* contribute to adaptation. Rather, they represent imperfections, whose functions were somehow preserved in spite of many "undesired" effects of gene conversion.

Birth-and-Death Evolution

The birth-and-death model for the evolution of gene families was proposed by Hughes and Nei (1989). In this model, new gene copies are produced by gene duplication. Some of the duplicate genes diverge functionally; others become pseudogenes owing to deleterious mutations or are deleted from the genome. The end result of this mode of evolution is a multigene family with a mixture of functional genes exhibiting varying degrees of similarity to one another plus a substantial number of pseudogenes interspersed in the mixture.

The birth-and-death model was initially put forward to explain the unusual pattern of evolution of the major histocompatibility genes in mammals (Hughes and Nei 1989; Klein et al. 1993; Nei et al. 1997). Subsequently, the model was used to explicate the evolution of such large and heterogeneous multigene families as the immunoglobulins and olfactory receptors, as well as that of highly conserved gene families such as the genes for histones and ubiquitins (Nei and Rooney 2005). Interestingly, even gene families consisting of only a few paralogs, such as the opsin and hemoglobin families, exhibit telltale signs of birth-and-death evolution (below).

Expansion and contraction of gene families

An important prediction of the birth-and-death process is that gene family size will vary among taxa as a result of differential birth and death of genes among different evolutionary lineages. Thus, an understanding of the evolutionary forces governing the birth-and-death process is predicated upon an accurate accounting of the number of births (duplications) and deaths in each lineage. This "bookkeeping" turns out to be anything but a trivial undertaking. Two computational methods have been employed for this type of analysis. One method requires a well-supported species tree and a detailed gene tree for each member of the multigene family. 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 (Zmasek and Eddy 2001). The second method uses maximum likelihood to infer family size at each internal node in the species phylogeny. It requires a priori knowledge of the number of gene copies in each family for each OTU, as well as estimates of divergence time between each pair of taxa (De Bie et al. 2006).

In **Figure 7.30**, a completely resolved phylogenetic tree for five *Saccharomyces* species is shown. There were 3,517 gene families shared by the five species. Of these, 1,254 (~37%) have changed in size across the tree. By inferring the most likely ancestral gene family sizes for all of these gene families, it was possible to deduce the number of changes in gene family size on all eight branches of the tree. Note that on each branch in the tree, the vast majority of gene family sizes remain static. Expansions outnumbered contractions on four of the eight branches, and contractions outnumbered expansions on the other four. Such evolutionary reconstructions can also be used to test for differences among lineages. For example, let us compare the numbers of expansions and contractions on the branches leading to *S. mikatae* and *S. cerevisiae* from their common ancestor, approximately 22 million years ago. On

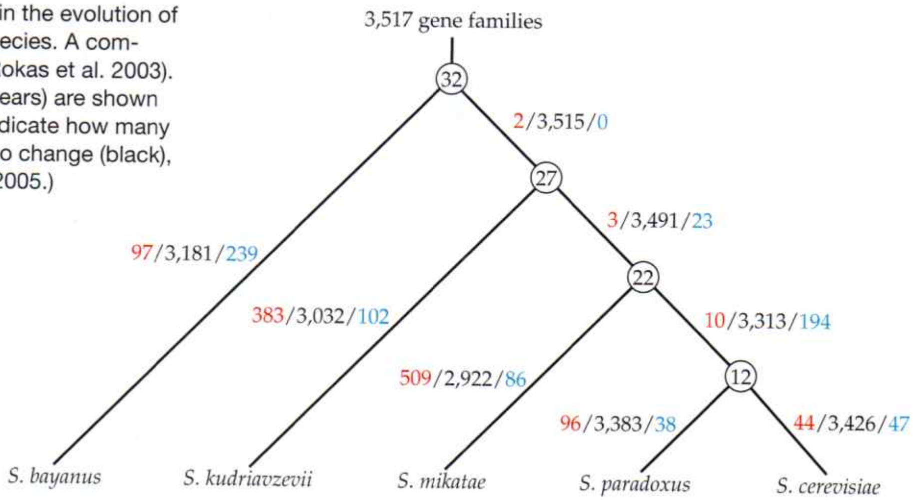


Figure 7.30 Expansions and contractions in the evolution of 3,517 gene families in five *Saccharomyces* species. A completely resolved phylogenetic tree is shown (Rokas et al. 2003). Estimates of divergence times (in millions of years) are shown in circles. The numbers along the branches indicate how many gene families experienced expansions (red), no change (black), or contractions (blue). (Data from Hahn et al. 2005.)

the lineage leading to *S. mikatae* there were 509 families that expanded and 86 families that contracted—a ratio of 6:1. In contradistinction, on the lineage leading to *S. cerevisiae* a smaller number of families changed their size, and the ratio of expanded families (54) to contracted ones (241) was inverted, 1:5. 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.

Rates of gene gain and loss are determined by an often difficult to disentangle interplay of mutation, fixation, and retention probabilities (Demuth and Hahn 2009). Analyses of gene family evolution in mammals revealed highly dissimilar rates of gene turnover across taxa. For example, the gene turnover rate in primates is nearly twice that in nonprimate mammals (0.0024 versus 0.0014 gains and losses per gene per million years). A further acceleration must have occurred in the great-ape lineage, such that humans and chimps gain and lose genes almost three times faster (0.0039 gains and losses per gene per million years) than the other mammals. In eukaryotes, new duplicates are "born" at the rate of 0.001–0.016 per gene per million years (Gu et al. 2002; Lynch and Conery 2003).

Rapid gene family expansion (amplification) in functionally important genes suggests adaptive scenarios in which natural selection favors additional copies either for increased dosage or for an increased arsenal of molecular weaponry. Particular cases illustrating this phenomenon have already been presented (p. 283). Here we look at general trends. In eukaryotes, gene amplification appears not to be common, as opposed to the situation in prokaryotes, where it is ubiquitous. The rarity of gene amplification in eukaryotes 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, among other effects, resistance to copper toxicity and growth under resource-limited conditions. As mentioned on page 293, independent amplifications of certain esterase genes are responsible for resistance to organophosphate pesticides (Field et al. 1988; Vontas et al. 2000), the most dramatic case being a 250-fold copy number increase in resistant strains of the mosquito *Culex pipiens* (Mouches et al. 1986).

In vertebrates, the evidence suggests a very limited role for gene amplification under positive selection. In mammals, less than 2% of all families display telltale signs of selective amplification. Gene families involved in immune defense, metabolism, cell signaling, chemoreception, and reproduction tend to amplify more frequently than would be expected by chance (see Demuth and Hahn 2009).

Examples of birth-and-death evolution

The birth-and-death process will be illustrated by the evolution of three gene families. Each was chosen to illuminate a different property or consequence of the process.

EVOLUTION OF THE OLFACTORY RECEPTOR GENE REPERTOIRE Olfactory receptors are G-coupled proteins that have seven α-helical transmembrane regions. Olfactory receptor genes are chiefly expressed in sensory neurons of the main olfactory epithelium in the nasal cavity. Vertebrates use different olfactory receptors and different combinations of olfactory receptors to detect many types of volatile or water-soluble chemicals. The number of functional olfactory receptor sequences in sequenced chordate genomes ranges from one in elephant shark to about 2,000 in the African elephant. The number of functional olfactory receptors is small in comparison to the number of known recognizable odorants, but olfactory receptors are known to function in a combinatorial manner, whereby a single receptor may detect multiple odorants, and a single odorant may be detected by multiple receptors (Malnik et al. 1999). Moreover, it was demonstrated that some olfactory receptors are "generalists" that bind to a variety of ligands, whereas others are "specialists" that are narrowly tuned to a small number of ligands.

In addition to intact olfactory receptor genes, the genomes of vertebrates contain numerous truncated genes and pseudogenes, both of which are presumably devoid of function. The percentage of nonfunctional olfactory receptor sequences ranges from 12% in zebrafish to 73% in tree shrew. The numbers of functional and nonfunctional olfactory receptor sequences in chordates are shown in **Figure 7.31**. (Olfactory receptors in invertebrates, including insects, nematodes, echinoderms, and mollusks, are not homologous to vertebrate olfactory genes. It therefore appears that genes encoding chemosensory receptors have evolved many times independently throughout animal evolution.)

As with other genes for G-coupled proteins, olfactory receptor genes do not have any introns in their coding regions. However, they often have introns and exons upstream of their coding regions. The noncoding exons can be alternatively spliced to generate multiple mRNA isoforms; however, this results in the same protein. Thus, it is not clear whether or not the alternative splicing of olfactory receptor genes has any biological significance. Olfactory receptor-coding genes constitute one of the largest gene families in tetrapods. For example, approximately 5% of the human proteome consists of olfactory receptors. Olfactory receptor genes form genomic clusters and are dispersed on many chromosomes. In the human genome, for instance, they are located on all chromosomes except 20 and Y.

Olfactory receptor gene loci make up one of the most genetically polymorphic regions in the human genome. Copy number variation is especially common, with many olfactory receptor loci being present in some individuals but not in others. Olfactory receptor gene coding regions also harbor a large number of single-nucleotide polymorphisms, some of which lead to inactivation of the functional gene and the creation of a segregating pseudogene. At least 35% of all human olfactory receptor genes exhibit copy number polymorphism or segregating pseudogenes. Therefore, the numbers of olfactory receptor genes vary enormously among individuals. This variation, in turn, causes olfactory perception to differ widely among individuals. "Specific anosmia" refers to the inability to perceive the odor of a specific substance in an individual having a generally good sense of smell. Specific anosmias are quite common (e.g., Gross 2007). In the United States, for instance, one in ten people cannot smell the extremely poisonous gas hydrogen cyanide, which for most people smells faintly like almonds; 12% cannot detect musky odors, which are common perfume ingredients; and one in 1,000 people cannot smell butyl mercaptan, the rancid issue of skunks, which is used as an additive to natural gas to enable its detection when it escapes or leaks from pipes.

The evolution of the olfactory receptor multigene family is characterized by an extremely rapid process of birth-and-death evolution, whereby genes are frequently added to the repertoire through gene duplication and neofunctionalization, and at the same time frequent gene losses occur by pseudogenization and deletion. The rapidity of the process can be deduced, for instance, by comparing the olfactory receptor repertoire of humans and chimpanzees, which have diverged from each other quite