CHAPTER 6

Large-Scale Gene and Ancient Genome Duplications

YVES VAN DE PEER AND AXEL MEYER

Duplications of genetic elements can occur by a variety of mechanisms and at different chromosomal and temporal scales. This chapter deals with an important subset of these, namely large-scale gene duplications (versus the small-scale events discussed in Chapter 5) and ancient duplications of whole genomes (versus more recent polyploidy in plants and animals, dealt with in Chapters 7 and 8, respectively). The emphasis in this case is on the techniques used to identify, date, and otherwise investigate such events, as illustrated by some key recent examples. As will be shown, analyses of different eukaryotes clearly indicate that significant portions of their genomes consist of duplicated gene loci, and that many of these gene duplicates have been formed by the duplication of chromosomal blocks and/or entire genomes. The timings of these events, in some cases dating back hundreds of millions of years, suggest that they have played an important role in influencing major patterns of evolutionary diversification.

The Evolution of the Genome, edited by TR Gregory

HISTORICAL PERSPECTIVES ON THE IMPORTANCE OF LARGE-SCALE DUPLICATIONS

As noted in Chapter 5, Ohno's (1970) book *Evolution by Gene Duplication* has become very influential in the field of genome research. This is a fairly recent phenomenon, with citations of the book tripling between 1990 and 2000, whereas it received only lukewarm reviews at the time of its publication (Wolfe, 2001). In the book Ohno (1970) made the case that not only gene duplications but doublings of entire genomes are the principal forces responsible for generating the genetic raw material necessary for increasing complexity during evolution. As he put it,

> Had evolution been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged. The creation of metazoans, vertebrates, and finally mammals from unicellular organisms would have been quite impossible, for such big leaps in evolution required the creation of new gene loci with previously nonexistent functions.

In other words, Ohno (1970) suggested that "natural selection merely modified, while redundancy created," meaning that gene and genome duplications allowed genes to diversify, take on novel functions, and bring about evolutionary innovation in general. Under Ohno's (1970) preferred interpretation of "neofunctionalization" (see Chapter 5), natural selection would be responsible for the fine-tuning of genes which had, through duplication, the chance to accumulate a sufficiently large number of otherwise "forbidden" mutations. The evidence for Ohno's hypothesis was based mainly on comparative measurements of DNA contents, karyotypic information, and some allozyme data. That is to say, Ohno's tenets were brought forth at a time where the documentation and quantification of genetic variation within populations and between species was largely restricted to scoring allelic variation in enzymes through starch gel electrophoresis and microscopic inspection of karyotypes. Methods to effectively measure genetic variation at the level of the gene or to even sequence DNA had still to be invented.

Like the case with small-scale duplications (see Chapter 5), Ohno (1970) was not the first to notice that the doubling of entire genomes could have been of major importance for evolution. In 1933, for example, Haldane argued that

> Duplications affecting only a few genes would confer only a slight advantage. But duplication of a large section, polysomy of a whole chromosome, or polyploidy, might confer a considerable advantage, provided it caused neither unbalance nor sterility. Whether this advantage is sufficient to be of evolutionary importance is not clear, but the possibility exists.

Haldane (1933) also described another possible mode of making rapid evolutionary jumps, namely by hybridization of the genomes of two different species. He noted that new species formed by hybridization showed heterosis (or "hybrid vigor"), with

increased fertility and stability, and therefore higher fitness, relative to their parent species.

As a matter of fact, one of the most important changes in agriculture over the past 50 years has been the improvement of many crops through the production of polyploid hybrids derived from the crossing of highly inbred lines. Whether through allopolyploidization (hybridization) or autopolyploidization (see later section), genome duplication is known to be a very common—perhaps even ubiquitous—occurrence in plant evolution (see Chapter 7). Indeed, many of the most important crop species are recent allopolyploids (e.g., wheat, oat, cotton, and coffee) or recent autopolyploids (e.g., alfalfa and potato), whereas others, such as cabbage, are ancient polyploids (Osborn *et al.*, 2003). Though not as common as in plants, polyploidy is also a widespread phenomenon in the animal kingdom. As discussed in detail in Chapter 8, many species (or sometimes higher taxa, like families) of fishes, amphibians, annelids, molluscs, crustaceans, insects, and other groups have been identified as recent or more ancient polyploids.

With the advent of large-scale genome sequencing, it has become possible to test some of the hypotheses put forth by Ohno and his predecessors by investigating, on a genomewide scale, whether large-scale duplication events of genes or even entire genomes have indeed been important over long evolutionary timescales.

MECHANISMS OF LARGE-SCALE DUPLICATION

In some groups, polyploids appear to form frequently and repeatedly, with a large percentage of species showing signs of recent polyploidization (see Chapters 7 and 8). There is also increasing evidence that many organisms are "paleopolyploids"—that is, ancient polyploids whose genome duplications have been masked by subsequent molecular and chromosomal evolution (see later section). The mechanisms involved in the initial duplications are thought to be the same as those occurring in more recent polyploidization events. Other mechanisms of large-scale duplications, above the level of individual genes but less than entire genomes, are also recognized. The most important of these are described briefly in the following sections. Additional details regarding mechanisms of polyploidization can be found in Chapters 7 and 8, whereas rediploidization, the evolutionary process in which a tetraploid species "decays" to become a diploid, is discussed in more detail by Wolfe (2001).

AUTOPOLYPLOIDY

Polyploidy can occur when an error during meiosis leads to the production of unreduced (i.e., diploid) gametes rather than haploid ones, as shown in Figure 6.1.

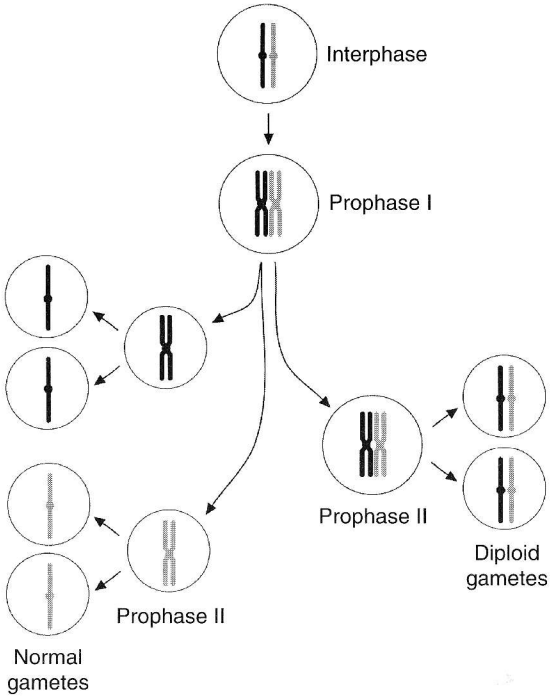

FIGURE 6.1 The basis of autopolyploidization. Autopolyploidization can occur when the pairs of homologous chromosomes have not separated into different nuclei during meiosis. The resulting gametes will be diploid rather than haploid. Based on Brown (1999), reproduced by permission (© BIOS Scientific Publishers).

If two diploid gametes fuse, an autotetraploid will be created whose nucleus contains four copies of each chromosome. Autopolyploids are often viable because each chromosome still has a homologous partner and can therefore form a bivalent during meiosis. This mechanism allows an autopolyploid to reproduce successfully, but prevents interbreeding with the original organism from which it was derived because a cross between a tetraploid and a diploid would give triploid offspring. Unlike tetraploids, triploids are very often sterile because one full set of its chromosomes lacks homologous partners to form the bivalents necessary for segregation (see Chapter 8 for more on the meiotic consequences of polyploidy).

Allopolyploidy

Polyploidy can also result from hybridization of two closely related species, leading to viable hybrids when the genomes are very similar, or to sterile hybrids when

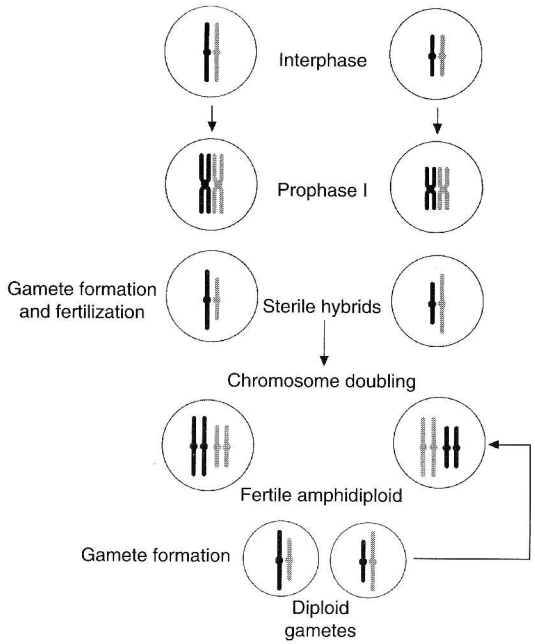

FIGURE 6.2 The basis of allopolyploidization. Allopolyploidy can result from hybridization of two closely related species, possibly leading to sterile hybrids, because chromosomes can not synapse (pair) during meiosis because they are not similar enough. However, when the newly combined genome undergoes a chromosomal doubling, two identical sets of chromosomes are available to pair during meiosis, and a fertile tetraploid is produced.

the chromosomes are insufficiently similar to synapse (pair) during meiosis. However, if the new combined genome undergoes a chromosomal doubling, two identical sets of chromosomes are available to pair during meiosis. As a result, a fertile tetraploid is produced (Fig. 6.2). For the most part, ancient polyploids are assumed to have been formed through allopolyploidy rather than autopolyploidy (e.g., Spring, 2003), although some semiancient polyploids such as the salmonid fishes are believed to be autopolyploid (see Chapter 8).

Aneuploidy

Aneuploids have a chromosome number that differs from an exact multiple of the haploid chromosome set. In such cases, a single chromosome is either lost or added from a normal diploid set of chromosomes. Duplication of individual chromosomes, described extensively for humans (and *Drosophila*), is either lethal

or results in serious genetic diseases such as Down syndrome, which is caused by the possession of three copies (trisomy) of Chromosome 21.

Block Duplications

Comparative studies have suggested that "block" (or "segmental") duplications—the duplication of large DNA segments—have been a continuing process during evolution. Block duplications are those in which many genes and their upstream regions are duplicated in a single event. However, for a long time it was unclear how such duplications could be generated, whether most of them occur intra- or interchromosomally, whether these tend to be found in direct or inverted orientation, and what sort of sequences are involved at the junctions. Recently, Koszul *et al.* (2004) have used a gene dosage assay for growth recovery in *Saccharomyces cerevisiae* to address these questions. They demonstrated that a majority of the revertant strains resulted from the spontaneous duplication of large DNA segments, both intra- and interchromosomal, ranging from 41 to 655 kilobases (kb) in size. In fact, in many cases dozens of genes were duplicated in a single event. The types of sequences at the breakpoints as well as their superposition with the replication map suggest that spontaneous large segmental duplications mainly result from replication accidents (Koszul *et al.*, 2004).

Tandem Duplications

Tandem duplications are duplications where the two copies of the duplicated region are located immediately adjacent to one another. The process of unequal recombination (or crossing-over) is widely viewed as responsible for the creation of tandem duplications. Well-known examples of gene complexes created by tandem duplications are the *Hox* gene clusters (discussed elsewhere in this chapter) and ribosomal RNA genes (see also Chapter 5).

HOW LARGE-SCALE GENE DUPLICATIONS ARE STUDIED

Identification of Block Duplications

The search for traces of ancient large-scale duplications has received much attention of late, with hypotheses about the number and age of polyploidization events in different eukaryotes a subject of much current debate (Wolfe, 2001; Durand, 2003). Evidence for large-scale gene- or entire genome-duplication events often

comes from the detection of block duplications. Identifying duplicated regions at the gene level is usually based on a within-genome comparison that aims to delineate regions of conserved gene content and order (i.e., "colinearity") in different parts of the genome. Disagreement can arise because the detection of colinear regions in genomes is not always straightforward (Gaut, 2001; Vandepoele *et al.*, 2002a).

In general, one attempts to identify a number of homologous gene pairs (typically referred to as "anchor points") in relatively close proximity to each other between two different segments in the genome, either on the same chromosome or on different chromosomes. When such a candidate colinear region has been detected, usually some sort of permutation test is performed in which a large number of randomized datasets is sampled in order to calculate the probability that the observed colinearity could have been generated by chance (Gaut, 2001; Simillion *et al.*, 2002). When the similarity between two genomic segments can be shown to be statistically significant, i.e., unlikely to be the result of chance, the conclusion is that the duplicated genes are the result of a single block duplication.

The statistics that determine colinearity thus depend on two factors: (1) the number of anchor points and (2) their distance from each other. These factors in turn usually depend on the number of "single" genes that interrupt colinearity. The tendency for a high level of gene loss, together with phenomena such as translocations and chromosomal rearrangements (see Chapter 9), often renders it very difficult to find statistically significant homologous regions in the genome, in particular when the duplication events are ancient. Fortunately, techniques are available for dealing with this issue, such as the map-based approach developed by Van de Peer and coworkers described in detail in the following section.

The Map-Based Approach

In order to detect chromosomal locations of colinear genes, it is necessary to search for regions that can be paired up because they contain sets of homologous genes. This requires a dataset containing all gene products and their absolute or relative positions in a genomic sequence. The map-based approach to analyzing such data involves only two parameters: (1) G, the "gap size," which specifies the maximum allowable number of intervening, nonhomologous genes between two homologous genes within a colinear segment, and (2) Q, the "diagonal quality" of the colinear regions (see later section).

To detect colinearity in two genomic fragments, a comparative search of all gene products (i.e., the amino acid sequences of the proteins) coded for in the relevant regions is performed using BLASTp (protein–protein Basic Local Alignment Search Tool) (Altschul *et al.*, 1997) (see www.ncbi.nlm.nih.gov/blast). The goal is to detect homologous gene products in the two regions, with two protein

sequences considered homologous when they share more than 30% sequence identity over an alignable region of at least 150 amino acids. Homology can still be determined when the matching sequences have an alignable region smaller than 150 amino acids, but this involves more complex analysis to compare the structure, and not just the sequence, of the proteins (using what is called the "homology-derived secondary structure prediction identity cut-off curve") (Rost, 1999).

Once obtained, the information on homologous genes is stored in a so-called "Gene Homology Matrix" (GHM), a hypothetical example of which is illustrated in Figure 6.3. In general terms, such a matrix consists of $m \times n$ elements, with m and n being the total number of genes on each genomic fragment. Pairs of homologous genes ("nonzero elements") in the matrix are identified by the coordinates (x, y). As shown in Figure 6.3A, colinear regions are represented as diagonal lines in the matrix, and tandem duplications are manifested as either horizontal or vertical lines, depending on which genomic segment has the additional copies. Inversions can be detected by looking at the organization of the entries. Gaps in

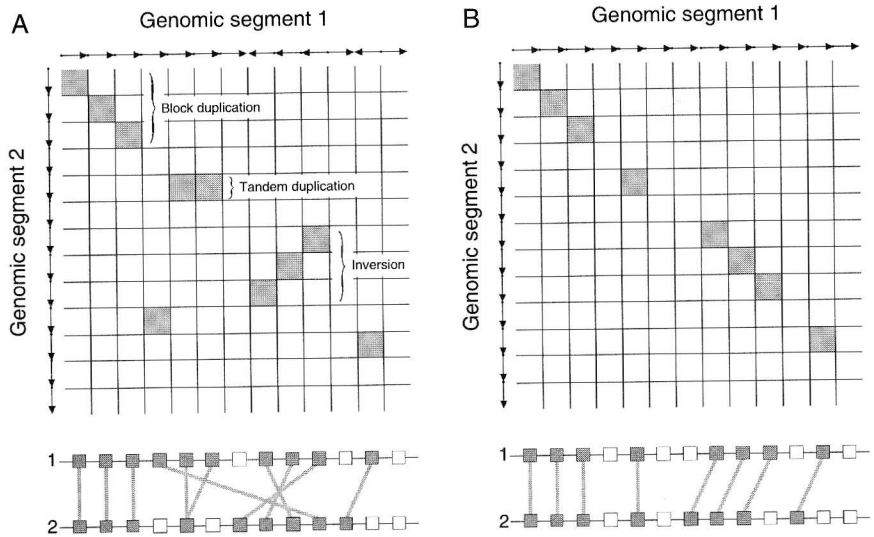

FIGURE 6.3 Hypothetical gene homology matrix (GHM). Each arrow on the axes of both segments represents a gene on the genomic segment. Gray cells illustrate homologous genes (called "anchor points"). **(A)** The original organization of all genes in their genomic context, with tandem duplications and inversions clearly visible. **(B)** The same gene homology matrix after tandem remapping and the removal of irrelevant (i.e., not part of a duplication) single data points by the ADHoRe algorithm. In addition, the small inverted colinear segment of three anchor points was restored to its original orientation in order to create a larger colinear region. See text for more details. From Vandepoele *et al.* (2004b), reproduced by permission (© Bentham Science Publishers Ltd.).

diagonal regions indicate insertions (through translocation, not duplication) or losses of genes in duplicated blocks.

After identification of the homologous genes, irrelevant data points need to be removed by a process referred to as "filtering" (Vandepoele *et al.*, 2002a). The fact that identifying colinearity effectively means finding diagonal series of elements in the matrix reduces the question to what is called a "clustering problem." This allows all elements that are too far away from other elements in the homology matrix to belong to a cluster to be removed during filtering. Next, the vertical and horizontal regions representing tandem duplications are deleted from the matrix. Specifically, these are remapped by collapsing all tandem duplications of a gene with the same orientation and within a distance G into a single element in the matrix. Tandem remapping makes it easier to detect diagonal regions, because then they are no longer interrupted by horizontal or vertical elements. The end result is a matrix in which a duplicated region now appears as a clear diagonal, as illustrated in Figure 6.3B.

In statistical terms, locating the diagonal regions in the matrix involves a special distance function that yields a shorter distance for points that are in diagonally close proximity than for points that are in horizontal or vertical proximity (Vandepoele *et al.*, 2002a). A generalized version of this is depicted in Figure 6.4A, and Figure 6.4B shows the application of this distance function to a hypothetical example. The actual clustering step is conceived as an iterative process, whereby the gap size is gradually increased until the final gap size (G) is reached. During each iteration, the gap size represents the maximum distance between two points in a cluster. Each time the process is repeated, new clusters can be formed and existing clusters can be extended. In the approach described here, by default the initial gap size is set to 3 and is then increased in 10 exponential steps until the final gap size has been reached.

Again, gap size is only one of the two parameters involved in this algorithm. The second is the "quality" of the clusters, meaning that it is important to only join genes to clusters that are assumed to have been created by the same duplication event. This is represented by the second parameter, Q, which determines the extent to which the elements of a cluster actually fit on a diagonal line. This "quality" parameter is estimated by calculating the coefficient of determination (r^2) by linear regression through the points in the clusters. Only clusters with a sufficiently high quality (i.e., higher than the cutoff Q) will be kept. Each addition of a potential gene duplicate to the diagonal line is tested, using the specific distance function described above, to determine the effects on the quality of the line. That is to say, each iteration of the algorithm involves a statistical test of whether the clusters can be enriched by adding single genes ("singletons") or joined with other clusters without badly affecting the cluster's diagonal properties.

Three conditions must be fulfilled for such additions or mergers to be accepted. First, the candidate singleton or cluster must be within a distance

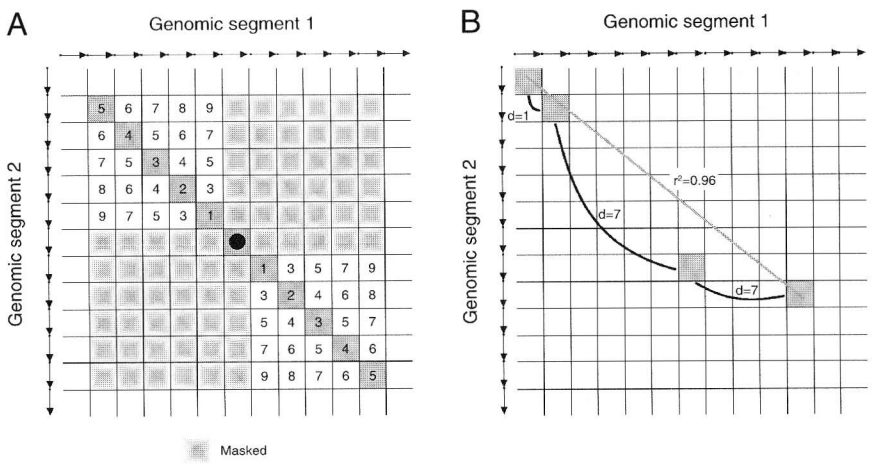

FIGURE 6.4 Application of the diagonal pseudo distance (DPD) function to the detection of elements with diagonal proximity in the gene homology matrix. (A) The DPD for a given cell in the matrix to the central black dot (anchor point). The diagonal pseudo distance is smaller for diagonally orientated elements (gray boxes) than for elements deviating from the diagonal. Shaded boxes represent elements (genes) with an infinite distance to the central dot, because these elements are unlikely to be part of the duplicated segment that contains the black dot. (B) The iterative clustering of elements for a colinear region with positive orientation (i.e., from top left to bottom right) in the homology matrix. All genes lie within a maximum gap distance G (for instance 30) of each other. The best-fit line and its coefficient of determination (r^2) show the quality of the cluster, which is clearly above the predefined Q value cutoff, here set to 0.9. As a result, all four homologous genes are considered to have arisen by a block duplication. From Vandepoele *et al.* (2004b), reproduced by permission (© Bentham Science Publishers Ltd.).

smaller than or equal to the current gap size in the iteration. Second, the candidate singleton must be positioned within the 99% confidence interval of the cluster (see Fig. 6.5). This confidence interval is computed by considering the best-fit line $y = ax + b$ through all the points in the cluster using the least-squares fit method. Usually, the points in the cluster show a certain degree of deviation from this line, which can be explained by two factors: (1) the error on the calculation of the constants a and b of the regression line, and (2) the error caused by the deviation of the point x_i, y_i from this line. Assuming this deviation is normally distributed, a confidence interval can be calculated that indicates the maximum deviation a candidate singleton can have from the best-fit line. If a singleton or cluster lies within these boundaries, then its effects on the r^2 of the diagonal line will be tested. If adding it does not cause the r^2 to fall below the cutoff (Q), it will be added to the cluster (Fig. 6.5B). The entire process is then repeated, using an increased gap size with this new cluster as the starting point. An example of the real-world application of such a process (using the ADHoRe software tool) to two fragments of the *Arabidopsis thaliana* genome is shown in Figure 6.6.

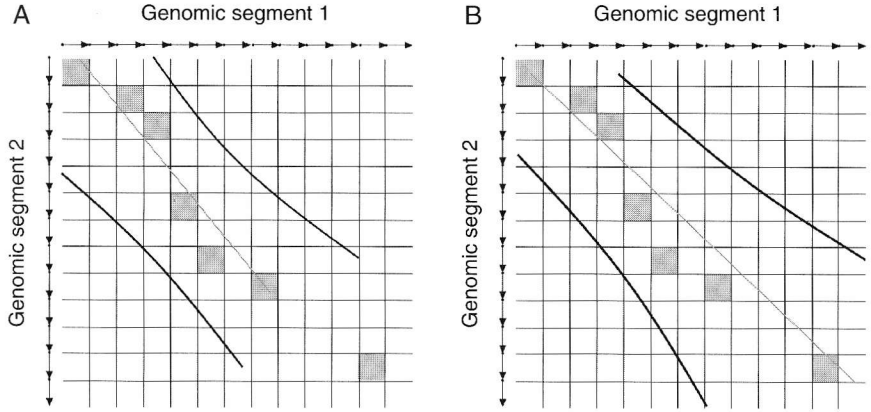

FIGURE 6.5 When adding genes to duplicated segments, it is assessed, using the specific diagonal pseudo distance (DPD) function (see Fig. 6.4), whether the clusters can be enriched with singletons (single genes) (A) or joined with other clusters without badly affecting the cluster's diagonal properties. To this end, the candidate singleton or cluster must be within a distance smaller than or equal to the gap size in the current iteration. Next, the candidate singleton must be positioned within the 99% confidence interval of the cluster. This confidence interval is computed by considering the best-fit line $y = ax + b$ through all the points in the cluster using the least-squares fit method. If these requirements are fulfilled, the segmental block duplication is extended (B).

Compiling a cluster (i.e., identifying a colinear region) is not the end of the procedure, because it is still necessary to remove any clusters that could have arisen by chance. This is accomplished with the use of a permutation test, by sampling a large number of reshuffled datasets and calculating the probability that a colinear region, characterized by a number of conserved genes and an average gap size, can be found by chance. When the similarity between two genomic segments can be shown to be statistically significant in this way, the conclusion is that both

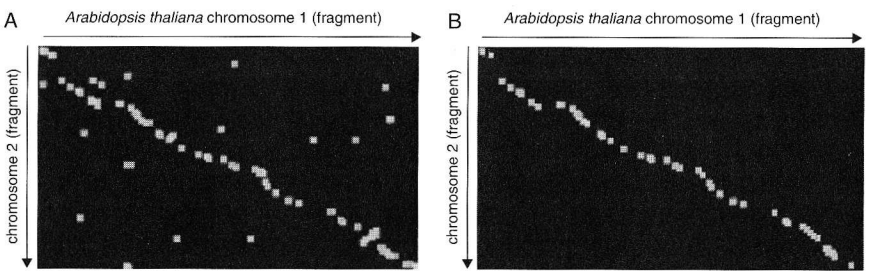

FIGURE 6.6 Example of the application of the ADHoRe software tool to two fragments of the *Arabidopsis thaliana* genome, (A) before and (B) after the filtering process (see Figs. 6.3–6.5).

segments are homologous and have originated by duplication. Permutation tests are very computer intensive, but recently novel, faster statistical methods have been developed to determine the statistical significance of putative homologous segments (Calabrese *et al.*, 2003; Simillion *et al.*, 2004). These methods are based on the observation that a cluster that was generated by chance generally contains fewer anchor points than a truly significant cluster, and that the average distance between these anchor points is also greater. In other words, the more anchor points a cluster contains and the closer these anchor points are located to each other on the diagonal of the GHM, the less likely it is that this cluster has been generated by chance.

Although the identification of block duplications is usually considered strong evidence for large-scale gene duplications, this is not a strict requirement. If many gene duplicates can be shown to have originated at about the same time in evolution, this could also be considered strong evidence that most of these paralogous genes have been created by one single event. Examples of such observations will be discussed later in this chapter.

Hidden Duplications, Ghost Duplications, and Multiplicons

In addition to the easily recognized "obvious" or "nonhidden" block duplications and tandem duplications (Fig. 6.3), there are also "hidden" and "ghost" duplications that are more difficult to identify (Fig. 6.7). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing both duplicated segments with each other, but only through comparison with a third segment within the genome. Consequently, hidden duplications are important when determining the actual number of duplication events that have occurred over time, as has been demonstrated previously for *Arabidopsis thaliana* (Simillion *et al.*, 2002). An example of such a hidden block duplication in *Arabidopsis* is presented in Figure 6.8.

Ghost duplications are defined as hidden duplications between different genomes. Two genomic segments in the same genome form a ghost duplication when their homology can only be inferred through comparison with the genome of another species (Vandepoele *et al.*, 2002b). In the case of *Arabidopsis* shown in Figure 6.8, for example, if Chromosome 2 proved to be derived from a different parental species than Chromosome 4, then the duplicated segments on Chromosomes 2 and 4 would form a ghost duplication.

As it turns out, a large number of chromosomal segments can often be identified as having been involved in multiple duplications. Such a group of homologous segments is referred to as a "multiplicon." Another way of displaying multiplicons is illustrated in Figure 6.9, which shows a network of colinearity between rice and *Arabidopsis*, including nonhidden, hidden, and ghost duplications.

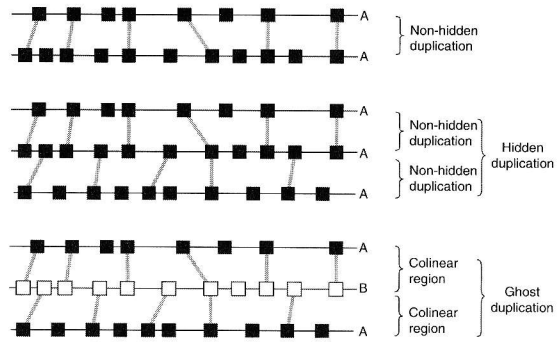

FIGURE 6.7 Schematic representation of nonhidden, hidden, and ghost duplications. Boxes represent the genes on chromosomal segments of genomes A and B, whereas connecting lines indicate the anchor points (i.e., homologous or duplicated genes). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing both duplicated segments, but only through comparison with a third segment from the same genome. Ghost duplications are hidden block duplications that can only be identified through colinearity with the same segment in a different genome. In contrast to hidden duplications, the identification of ghost duplications increases the fraction of the genome involved in a duplication event. From Vandepoele *et al.* (2003), reproduced by permission (© American Society of Plant Biologists).

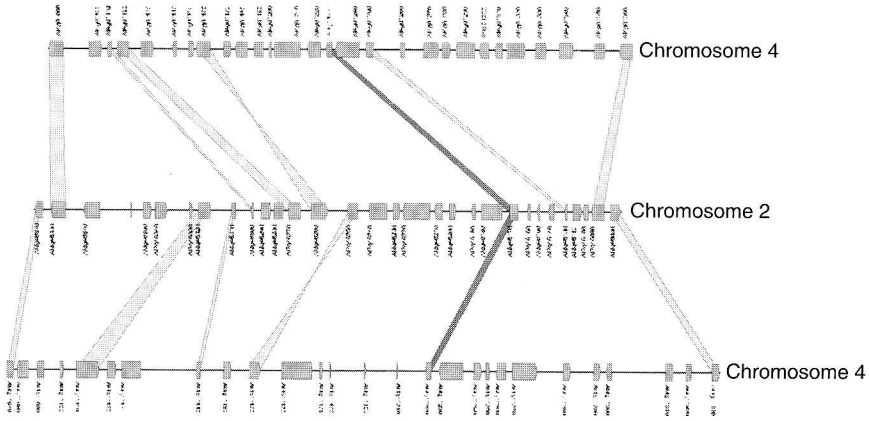

FIGURE 6.8 Example of a multiplicon in *Arabidopsis thaliana*. No duplication can be observed between the two segments on Chromosome 4, because these have only one homologous gene in common (dark gray band). However, both segments still share several, but different, homologous genes with a segment on Chromosome 2. Therefore, both segments on Chromosome 4 form a hidden duplication. If Chromosomes 2 and 4 were found to be derived from two different species, then this would constitute a ghost duplication.

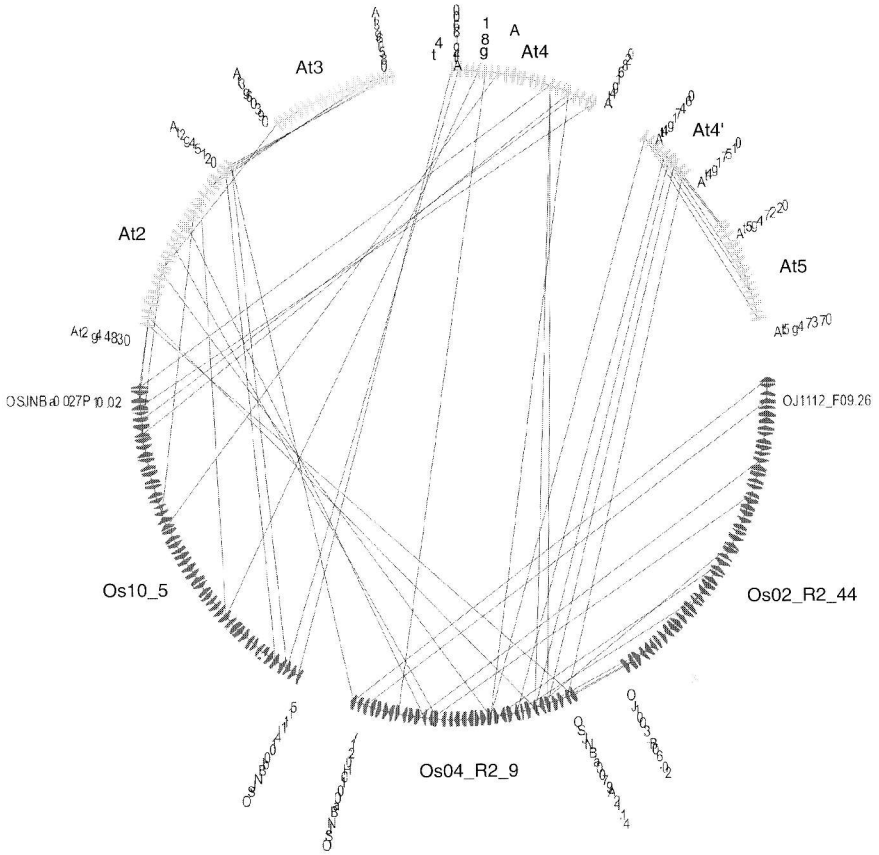

FIGURE 6.9 Set of homologous chromosomal segments (multiplicon) of *Arabidopsis thaliana* (At) and rice (*Oryza sativa*, Os). Arrows represent the genes on the chromosomal segments, whereas connecting lines indicate the anchor points (i.e., homologous or duplicated genes) that are part of a significant colinear region determined by the ADHoRe algorithm. For each genomic segment, the names of the two genes delineating the segment are shown. Chromosomal segments of rice and *Arabidopsis* are shown in dark and light gray, respectively. By considering the colinearity between *Arabidopsis* and rice, a set of at first sight unrelated *Arabidopsis* segments can be joined into a multiplicon with multiplication Level 5, confirming the three duplication events in *Arabidopsis* described earlier (Simillion *et al.*, 2002). Conversely, colinearity between rice and *Arabidopsis* reveals that all three rice segments are linked with each other by two duplication events.

Genomic Profiles: An Extension to the Map-Based Approach

Although considering transitive homologies such as hidden and ghost duplications allows the identification of many previously undetectable homologous genomic segments, it still requires that these show significant colinearity with at least one other homologous segment. However, it is possible that, within a given multiplicon, one or more segments have diverged so much from the others in gene content and order that they no longer show clear colinearity with any of the other segments. Unfortunately, such segments in the "twilight zone" of genomic homology cannot be detected with any of the currently available methods. New software is being developed (e.g., by Van de Peer and colleagues) to uncover chromosomal segments that are homologous (with respect to having common ancestry) to others but can no longer be identified as such because of extreme gene loss. This is done by aligning clearly colinear segments and using this alignment as a "genomic profile" that combines gene content and order information from multiple segments to detect these heavily degenerated homology relationships (see Fig. 6.10).

After the initial detection of a "Level 2" multiplicon (i.e., a pair of homologous chromosomal segments) with the basic ADHoRe algorithm, an alignment of the two segments that form this multiplicon can be created where the anchor points of the multiplicon are positioned in the same columns. Using this alignment as a "profile," a new type of homology matrix can be constructed in which the gene products of a segment are compared to the gene products of the profile. Once this new GHM is constructed, it is subjected to the basic ADHoRe algorithm, which involves the same statistical validation procedures to detect clusters of anchor points. This time, however, new significant clusters will not reveal homology between two individual segments, but rather between the two segments inside the profile (i.e., the initial Level 2 multiplicon) and a third segment. Because this type of GHM combines gene content and order information of the different segments in the profile, it is possible to detect homology relationships with a third segment that could not be recognized by directly comparing any of the segments of the multiplicon individually with this third segment. If such a third segment is detected, it is added to the multiplicon, thereby increasing its multiplication level, and the corresponding profile is updated by aligning the new segment to it. The entire detection process is then repeated with the newly obtained profile.

By constructing genomic profiles that combine gene content and order information from multiple homologous segments, it becomes possible to detect heavily degenerated homology relationships between segments that no longer show significant colinearity with any of the segments contained in the profile. The strength of this approach is clearly illustrated by the substantial increase in multiplicons it generates in *Arabidopsis* as compared with the traditional approach; indeed, multiplications of Level 5 or greater may be observed in this way (see Simillion *et al.*, 2004).

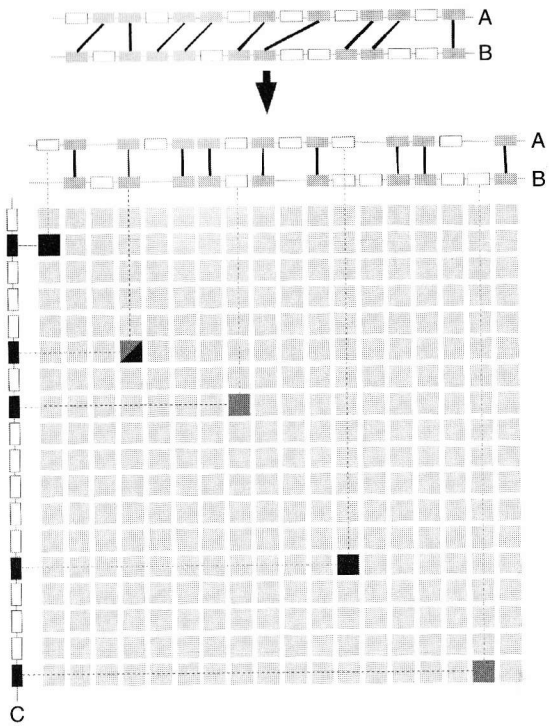

FIGURE 6.10 Detection of homology through a genomic profile. The upper section shows an initially detected Level 2 multiplicon (a pair of homologous chromosomal segments). The gray boxes connected by black lines represent pairs of homologous genes (anchor points) between the two segments. The lower section shows the construction of a homology matrix using this multiplicon as a profile. To accomplish this, the multiplicon is first aligned by inserting gaps at the proper positions (depicted by empty spaces in the alignment). The homology matrix can now be constructed by comparing this profile with the genes of a chromosomal segment C (shown on the left of the matrix). Anchor points in the matrix are detected whenever a gene of this chromosomal segment belongs to the same gene family as one of the genes in any of the segments in the profile. The black squares represent homologs between segments A and C, and the dark gray between B and C. The black/dark-gray square denotes a gene that has a homolog on both segments A and B. Combining segments A and B in a profile thus results in five anchor points with segment C, whereas the individual segments A and B only have three anchor points with segment C, which might be too few to detect statistically significant homology.

DATING DUPLICATION EVENTS

Several methods are commonly being used to date gene duplication events, the most notable of which are (1) absolute dating based on third codon or synonymous substitution rates, (2) absolute dating based on nonsynonymous substitution rates or protein-based distances, and (3) relative and absolute dating by the construction and analysis of phylogenetic trees. These will be discussed in turn.

ABSOLUTE DATING BASED ON SYNONYMOUS SUBSTITUTIONS

Because most substitutions in third-codon positions do not result in amino acid replacements (Fig. 6.11), the rate of fixation of these substitutions is expected to be relatively constant in different protein-coding genes (Nei and Kumar, 2000) and to reflect the overall mutation rate (Hughes, 1999a). Time of divergence (T) can be calculated from this as $T = K_S/2\lambda$, where K_S is the fraction of synonymous substitutions per synonymous site and λ is the mean rate of synonymous substitution (Nei and Kumar, 2000). The value for λ differs for various organisms; in *Arabidopsis*, for instance, the estimate is 6.1 synonymous substitutions per 10^9 years, whereas for mammals it is considered to be about 2.5 substitutions per 10^9 years (Lynch and Conery, 2000).

Although silent substitutions have been used extensively to compute duplication events, there is one important caveat, namely that dating based on such substitutions can only be applied when K_S is less than 1. Higher values of K_S point to saturation of synonymous sites and should therefore be used with great caution when drawing any conclusions regarding the date of duplication events. There are different ways to compute the number of synonymous substitutions per synonymous site, depending on which method is used to correct for multiple mutations at these sites. For example, the NTALIGN program in the NTDIFFS software package (Conery and Lynch, 2001) first aligns the DNA sequence of two mRNAs based on their corresponding protein alignment and then calculates K_S by the method of Li (1993). Nei and Gojobori (1986) and Yang and Nielsen (2000) have proposed two alternative methods to compute K_S, both of which are implemented in the PAML phylogenetic analysis package (Yang, 1997).

PROTEIN-BASED DISTANCES

Although protein-based distances are known to vary considerably among proteins (Easteal and Collet, 1994) because of different functional constraints, several attempts have been made to use such distances to date duplication events. For example, Vision *et al.* (2000) have used amino acid replacement rates (K_A) to date

```
 M   A   L   A   F   D   E   F   G   R   P   F   I   I   L
ATG GCT TTG GCT TTC GAT GAG TTT GGC CGG CCG TTC ATT ATA CTA Duplicate α
ATG GCG CTG GCG TTC GAT GAG TTC GGG CGT CCG TTC ATT ATA CTG Duplicate β
```

FIGURE 6.11 Silent substitutions, indicated in bold, mostly occurring at third codon positions, do not lead to amino acid replacement and are therefore regarded as "neutral," and assumed to follow a clocklike behavior.

block duplication events in *Arabidopsis*. These authors assumed that, whereas the mutation rate of different proteins may vary considerably, the overall distribution of amino acid substitution rates is the same throughout the genome. If that assumption were valid, then any contemporaneously duplicated block containing several homologous pairs would provide a more or less independent sample of the distribution. Furthermore, the average values of K_A for blocks duplicated at the same time must necessarily be much less variable around the true mean than the individual protein values themselves. Unfortunately, there is some evidence from other organisms that rates of protein evolution vary systematically in different regions of the genome. However, for that phenomenon to create problems with dating based on the block averages, the variation among regions would have to be on the same scale as the differences between duplicated blocks of different age classes, and to co-vary among the chromosome pairs in each block (T. Vision, personal communication).

That said, it has been shown that protein distances are not very reliable for dating duplicated blocks containing heterogeneous classes of proteins. For example, different block duplications in *Arabidopsis* estimated to be of similar age based on mean protein distance (Vision *et al.*, 2000) actually turned out to be very heterogeneous in age when compared to dating based on synonymous substitution rates (Raes *et al.*, 2003). The reason is that duplicated blocks that contain a larger fraction of fast-evolving genes will have a relatively high mean protein distance between the paralogous regions and appear older than they actually are. It would therefore seem that the use of synonymous and, consequently, neutral substitutions for evolutionary distance calculations is the more reliable way of estimating duplication events, unless there is no alternative because the duplications are too old.

Dating by Phylogenetic Means

Another way of dating duplication events is by mapping them onto phylogenetic trees. In relative terms, this approach allows a determination of whether duplications have occurred prior to or after a speciation event. For example, in Figure 6.12A, the gene has been duplicated prior to the divergence of zebrafish and pufferfish (~ 150 million years ago), whereas the gene duplications in Figure 6.12B are younger, and have occurred independently in zebrafish and pufferfish after their divergence.

If the timing of a speciation event is known with confidence, gene trees can also be used to infer absolute dates. This is usually performed by the construction of linearized trees (Takezaki *et al.*, 1995), which assumes equal rates of evolution in different lineages of the tree—that is, a molecular clock (see Chapter 9). In order to create such linearized trees, relative rate and branch length tests for rate

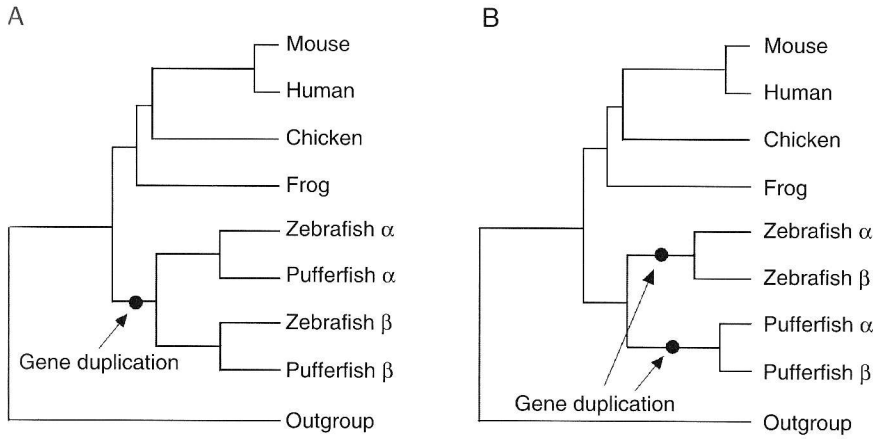

FIGURE 6.12 Relative dating of duplication events by phylogenetic means. Different scenarios—and expected inferred tree topologies—are shown to explain the presence of more genes in fishes. (A) Duplicated fish genes resulting from a gene/genome duplication that preceded the divergence of zebrafish and pufferfish. (B) Duplicated genes formed by independent gene duplications.

heterogeneity are usually applied to these trees to check for deviations from the assumption of a constant molecular clock. Faster or more slowly evolving sequences are then removed so that the dataset contains only sequences evolving at a similar rate. By comparing the divergences of duplicated genes with a fixed calibration point—that is, the date of a particular evolutionary event, such as the divergence between fishes and land vertebrates—the absolute date of origin of paralogous genes can be inferred.

PUTTING THEORY INTO PRACTICE: EVIDENCE FOR LARGE-SCALE GENE DUPLICATION EVENTS

Although there is evidence that individual gene duplications occur frequently and are actually part of a continual process (Lynch and Conery, 2000; Gu *et al.*, 2002) (see Chapter 5), more and more genomic data seem to suggest that many gene duplicates have arisen during major large-scale duplication events. Indeed, ancient duplications of entire genomes have now been documented for members of the three best-studied eukaryotic kingdoms. The first strong evidence for an ancient polyploidy event in eukaryotes came from the yeast *Saccharomyces cerevisiae*. Based on a genomewide analysis, it was postulated that the entire yeast genome had duplicated about 100 million years ago (Wolfe and Shields, 1997), and that as a result, approximately 25% of the yeast genome still consists of duplicated genes

(Seoighe and Wolfe, 1999). Recently, the genome duplication in yeast has been confirmed through comparative analysis with closely related species (Dietrich *et al.*, 2004; Dujon *et al.*, 2004; Kellis *et al.*, 2004). As described in the following sections, some intriguing examples are now also known from animals and plants.

1R/2R: Genome Duplications in Vertebrates

In *Evolution by Gene Duplication*, Ohno (1970) argued that large-scale gene duplication occurred during the evolution of early vertebrates. Although based on rather inaccurate indicators of genome complexity, such as genome size (see Chapter 1) and isozyme patterns, Ohno proposed that two rounds of genome duplications had occurred in the evolutionary past of early vertebrates, one on the shared lineage leading to both cephalochordates and vertebrates, and a second in the fish or amphibian lineage (see also Furlong and Holland, 2002) (Fig. 6.13).

The advent of DNA sequence–based analysis provided more reliable evidence for the hypothesis of two rounds of large-scale gene duplications in the early vertebrates. A prime example of this is the analysis of *Hox* genes (Holland *et al.*, 1994). *Hox* genes encode DNA-binding proteins that specify cell fate along the anterior–posterior axis of bilaterian animal embryos, and occur in one or more clusters of up to 13 genes per cluster (reviewed in Gehring, 1998). The observation that protostome invertebrates, as well as the deuterostome cephalochordate *Branchiostoma lanceolatum* (commonly called "Amphioxus"), possess a single

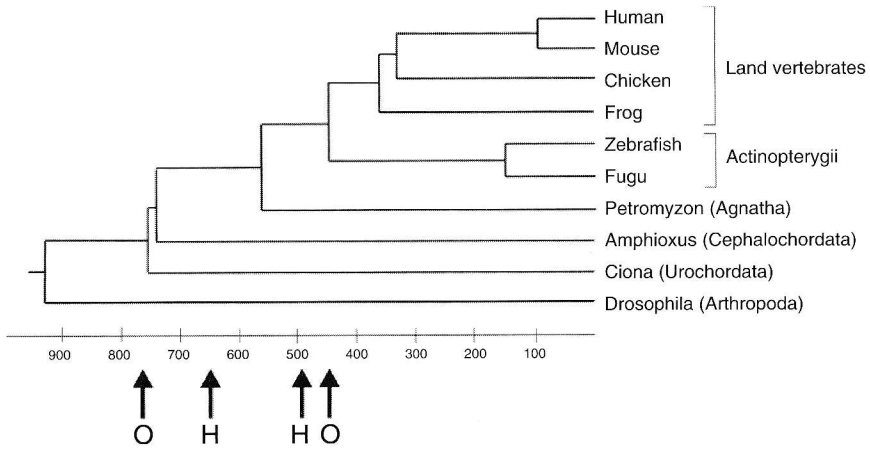

FIGURE 6.13 Phylogenetic tree of major vertebrate groups and their time of divergence. Arrows indicate presumed genome duplications according to (O) Ohno (1970) and (H) Holland *et al.* (1994).

Hox cluster, whereas the lobe-finned fishes (coelacanth and lungfishes), amphibians, reptiles, birds, and mammals have four clusters (Holland and Garcia-Fernandez, 1996; Holland, 1997; Larhammer *et al.*, 2002), supports the hypothesis of two rounds (2R) of entire genome duplications early in vertebrate evolution. Holland *et al.* (1994) proposed that a first duplication occurred on the vertebrate lineage after the divergence of the cephalochordates, and a second one after the divergence of the jawless vertebrates (Fig. 6.13). Although some support can be found for this hypothesis (see, for example, Escriva *et al.*, 2002), the recent discovery of three, and most probably four *Hox* clusters in the lamprey *Petromyzon marinus* suggests that the two rounds of (*Hox* cluster) duplications occurred before the divergence of lampreys and hagfishes (Irvine *et al.*, 2002; Vandepoele *et al.*, 2004a). By contrast, some other authors assume an independent duplication history of the lamprey *Hox* clusters, and therefore do not consider this evidence for two rounds of large-scale gene duplication events prior to the divergence of lampreys and hagfishes from the other vertebrates (Fried *et al.*, 2003).

Spring (1997) uncovered an average of three paralogs in humans for each of 52 *Drosophila* genes and proposed that the additional human genes were produced during two allopolyploidization events in the early vertebrate lineage. The presence in four copies of various segments in vertebrate genomes has been reported in subsequent studies, which likewise is suggestive of two large-scale duplications (Abi-Rached *et al.*, 2002; Lundin *et al.*, 2003). Additional support for 1R or 2R of genome duplication comes from the detection and dating of duplicated blocks in the human genome and from large-scale phylogenetic analyses of gene families (Abi-Rached *et al.*, 2002; Gu *et al.*, 2002). Recently, McLysaght *et al.* (2002) described an extensive gene duplication during early chordate evolution. They suggested that at least one (maybe two) round(s) of polyploidization occurred in the early history of vertebrates, and concluded that humans, like yeast and *Arabidopsis* (see later section), are ancient polyploids. Gu *et al.* (2002) showed that both large- and small-scale duplications are required to explain the age distribution of duplicated human gene families.

Although a consensus seems to be emerging that large-scale gene or even entire genome duplication events have occurred in the evolution of early vertebrates, rediploidization and degeneration of duplicate genes generally makes strong evidence in support of 2R hard to find. As a consequence, the 1R/2R hypothesis of vertebrate genome evolution is still hotly debated, with opinions ranging from strongly in favor (e.g., Holland *et al.*, 1994; Furlong and Holland, 2002; Larhammer *et al.*, 2002; Panopoulou *et al.*, 2003; Spring, 2003; Vandepoele *et al.*, 2004a) to highly skeptical (e.g., Hughes *et al.*, 2001; Martin, 2001; Friedman and Hughes, 2003).

Much of this confusion may stem from the nature of the duplication events themselves, in particular their timing relative to each other. For example, some advocates of the 2R hypothesis believe that the two rounds of genome duplications

occurred in very short succession (Larhammer *et al.*, 2003). This would explain why it is generally hard to infer phylogenetic trees of the form ((A,B)(C,D)) using gene duplicates, which in principle should be easy to do if two tetraploidization events had occurred (Skrabanek and Wolfe, 1998; Hughes, 1999b; Martin, 2001). If both genome doublings indeed occurred almost contemporaneously, it is not surprising that they cannot easily be distinguished based on age differences between genes or the topology of gene family trees. However, as more large-scale genome sequence data become available, it should be possible to improve the resolution of such analyses and perhaps to answer this question conclusively.

3R: An Additional Round of Genome Duplication in Teleost Fishes

A few years ago, it was proposed that an additional (3R) genome duplication had occurred in ray-finned fishes (Aparicio *et al.*, 1997; Amores *et al.*, 1998; Wittbrodt *et al.*, 1998). As with the proposed duplication event(s) shared by all vertebrates, the first indications for a fish-specific genome duplication came from studies of *Hox* genes. Extra *Hox* gene clusters have been discovered in the zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), the African cichlid *Oreochromis niloticus*, and the pufferfish *Takifugu rubripes*. The observation that such distantly related species all share this feature suggested the occurrence of an additional genome duplication event in the ray-finned fish lineage (Actinopterygii) before the divergence of most teleost species (Amores *et al.*, 1998; Wittbrodt *et al.*, 1998; Meyer and Schartl, 1999; Naruse *et al.*, 2000; Málaga-Trillo and Meyer, 2001).

More recent comparative genomic studies have turned up many more genes and gene clusters for which two copies exist in fishes but not in other vertebrates (e.g., Postlethwait *et al.*, 2000; Woods *et al.*, 2000; Robinson-Rechavi *et al.*, 2001a; Taylor *et al.*, 2001, 2003; Van de Peer *et al.*, 2001). The findings that different paralogous pairs appear to have originated at about the same time (Taylor *et al.*, 2001), that different fish species seem to share ancient gene duplications (Taylor *et al.*, 2003), and that different paralogs are found on different linkage groups in the same order (i.e., show synteny) with other duplicated genes (Gates *et al.*, 1999; Postlethwait *et al.*, 2000; Woods *et al.*, 2000), all support the hypothesis that these genes arose through a complete genome duplication event. However, it bears noting that some authors have argued that an ancestral whole-genome duplication event was not responsible for the abundance of duplicated fish genes. For example, Robinson-Rechavi *et al.* (2001a,b) counted orthologous genes in fishes and mice and, where extra genes were found in fishes, compared the number of gene duplications occurring in a single fish lineage with that shared by more than one lineage. Most mouse genes surveyed were also found as single copies in fishes. Duplicated fish genes were detected, but most were interpreted

as the products of lineage-specific duplication events in fishes and not as a single ancient duplication event.

In order to find further evidence for or against large-scale gene duplication events in early vertebrate evolution, Vandepoele *et al.* (2004a) recently analyzed the complete genomes of the pufferfish *Takifugu rubripes* ("Fugu") and human. Phylogenetic trees were constructed for all (i.e., 3077) gene families containing two to 10 duplicated Fugu genes, and relative dating of duplication events was performed to test whether gene duplications occurred before (1R/2R) or after (3R) the divergence of the lineages that led to ray-finned fishes and land vertebrates (Fig. 6.14). This analysis showed that most paralogous genes in pufferfish are the result of at least two, probably three, complete genome duplications.

Absolute dating of duplication events was performed through the inference from linearized trees (Takezaki *et al.*, 1995). In these linearized trees—where branch-length is drawn directly proportional to time—the split between ray-finned fishes and land vertebrates (dated at 450 million years ago) (Carroll, 1988; Benton, 1990, Zhu *et al.*, 1999) was used as a calibration point for the dating of gene duplication events. The removal of trees with insufficient statistical support left 595 nodes, based on the analysis of 488 gene families, for which an absolute date could be inferred. Combining the results of relative and absolute dating, these 565 duplications could be subdivided into 166 3R and 399 1R/2R duplications (Figs. 6.14 and 6.15).

Put another way, these results indicate that a major fraction (30%) of the Fugu paralogs is younger than the split between ray-finned fishes and land vertebrates, probably arising somewhere between 225 and 425 million years ago. The most plausible and parsimonious explanation for this observation would be a large-scale

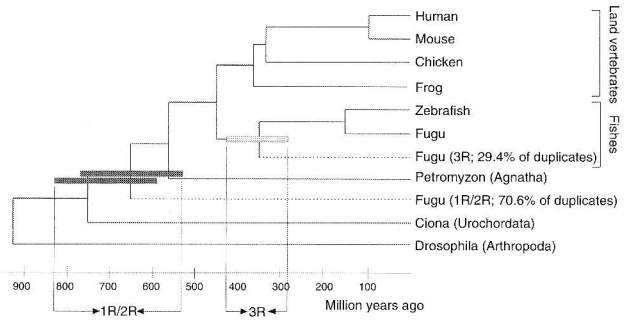

FIGURE 6.14 Phylogenetic tree of major vertebrate groups and superimposed Fugu gene duplication events. Black and gray bars denote large-scale gene duplication events observed in the Fugu genome based on absolute and relative dating and the detection of segmental duplications (see text for details). The time of divergence for the lamprey *Petromyzon*, as a representative of the Agnatha, was taken from Shu *et al.* (1999). From Vandepoele *et al.* (2004a), reproduced by permission (© National Academy of Sciences USA).

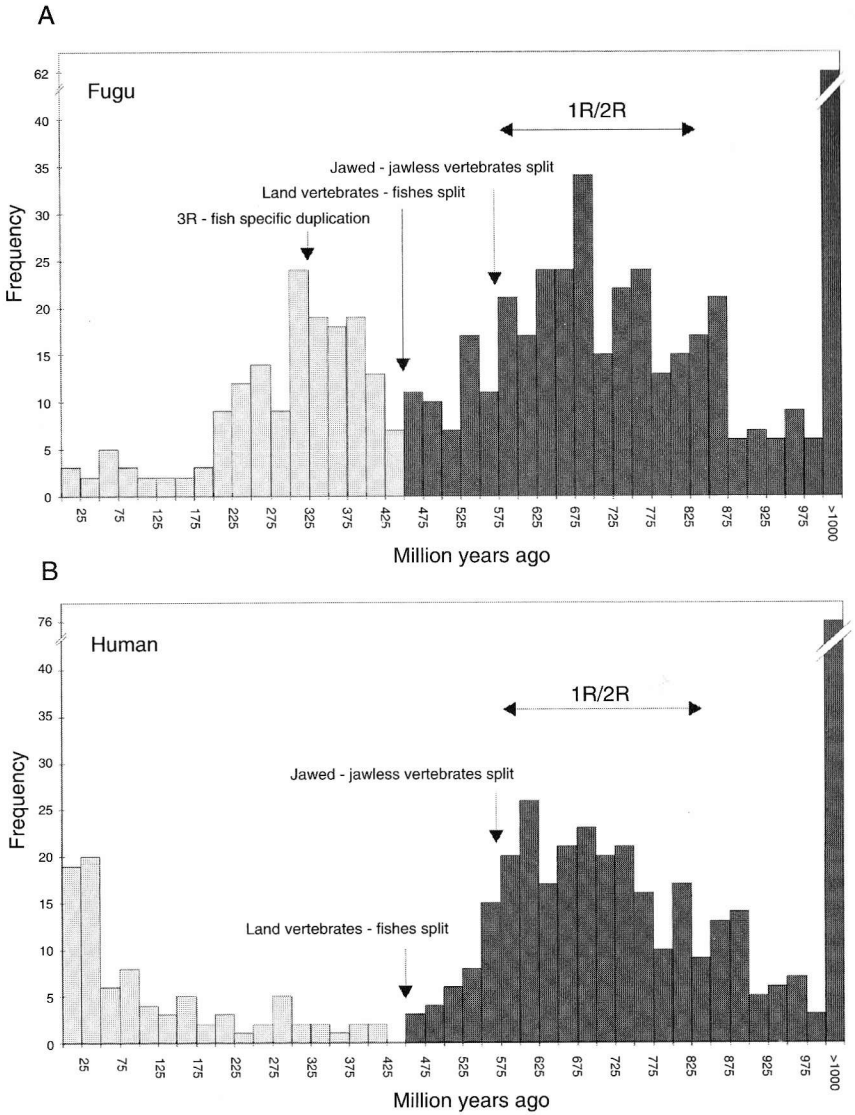

FIGURE 6.15 Age distribution of duplicated genes in the (A) Fugu (pufferfish) and (B) human genomes. Dark bins correspond to duplications prior to the divergence of ray-finned fishes and land vertebrates; light bins correspond to duplications after the split between ray-finned fishes and land vertebrates (see text for details). From Vandepoele *et al.* (2004a), reproduced by permission (© National Academy of Sciences USA).

gene or entire genome duplication. To test whether the sudden increase in the number of duplicated genes in the Fugu genome is the result of an entire genome duplication rather than an increased rate of independent tandem duplication events, Vandepoele *et al.* (2004a) investigated whether these duplicated genes appear in duplicated blocks on chromosomes (again, the identification of duplicated blocks is usually considered strong evidence for large-scale gene duplication events).

Using the map-based approach outlined above, Vandepoele *et al.* (2004a) identified statistically significant regions of microcolinearity (showing the same gene content and gene order) within the complete Fugu genome. All genes within such a region are presumed to have been duplicated at the same time, and hence to be of identical age, because it is unlikely that these colinear regions would be created independently on different chromosomes. By applying the ADHoRe algorithm to scaffolds of the available pufferfish genome sequence, and using phylogenetic methods to date the duplicated blocks so identified, Vandepoele *et al.* (2004a) were able to conclude that the 3R blocks of duplicated genes all arose at approximately the same time, namely about 320 million years ago, with a standard deviation of 67 million years.

Of course, it might be argued that a standard deviation of 67 million years is rather large and could indicate the occurrence of several independent block duplications rather than a single genome duplication event. However, when using an absolute dating approach, such a variance on estimated duplication times is to be expected even when the duplicates are of the same age (Vandepoele *et al.*, 2004a). In particular, within the same block duplication, homologous genes that have been duplicated at the same time can exhibit a considerable difference in estimated duplication time owing to deviations of the molecular clock. The simultaneous duplication for these genes is supported statistically (Vandepoele *et al.*, 2004a), even with the very fragmented nature of the Fugu scaffold dataset used in the analysis. In fact, the number of duplicated blocks is probably much higher, and is expected to rise considerably once better assemblies of the Fugu genome become available. This suggests that the wide distribution of duplicated Fugu genes already observed is in perfect agreement with the hypothesis of a single complete genome duplication event. Overall, such considerations provide very strong support for a complete genome duplication event in the early stages of fish evolution, predating the origin of most modern ray-finned fish species that are believed to have (started to) diverge(d) from each other more than 200 million years ago.

Additional evidence for a fish-specific 3R duplication event comes from analyzing nonfish genomes. Using the same methods as for Fugu, Vandepoele *et al.* (2004a) performed an analysis of gene duplicates in the human genome. Of the 447 duplication events identified in the human genome sequence, absolute dating suggested that 360 can be attributed to 1R/2R whereas 87 were specific to humans (see Fig. 6.15B). The distribution of inferred ages of duplicated genes shows a similar increase in the number of duplication events around 675 million

years ago, as observed in Fugu. Not only does this support the 1R/2R hypothesis, but it also confirms the expectation that no evidence of the hypothesized fish-specific 3R genome duplication event is found in the human genome (Fig. 6.15B).

To summarize, the relative and absolute dating of hundreds of gene families, together with the detection of many duplicated blocks that have originated at about the same time, provides strong support for the hypothesis of a fish-specific genome duplication ~320 million years ago that was not experienced in the lineage of vertebrates leading to humans. This 3R genome duplication event accounts for the large majority of gene duplicates found in the Fugu genome, in contrast to the situation in the human genome, where many more recent tandem and segmental duplication events account for the majority of duplicated genes (Bailey *et al.*, 2002) (Fig. 6.15). Most of the remaining paralogs seem to have been created by one or two much older large-scale duplication events, predating the split between ray-finned fishes and terrestrial vertebrates. Indeed, using the fish-specific genome duplication as a benchmark, and assuming equal rates of gene loss throughout vertebrate evolution, two genome duplications rather than one seem to have occurred—as proposed by Ohno in 1970.

Ancient Genome Duplications in Plants

As discussed in Chapter 7, estimates of the prevalence of polyploidy in flowering plants have been increasing over time, beginning at about 30–50% in the 1930s and 1950s, to 70–80% in the 1980s and 1990s. Today, it is becoming more common to suggest that 100% of angiosperms have polyploidy in their ancestry. Much of this new view is based on the discovery of ancient polyploidy even in plants in which it was not at all expected.

As a most notable example, although initial sequencing of the tiny genome of *Arabidopsis thaliana* revealed numerous duplicated segments (Paterson *et al.*, 1996; Lin *et al.*, 1999; Mayer *et al.*, 1999; Terryn *et al.*,1999), this plant was long believed to be a clear example of a diploid organism. However, after bacterial artificial chromosome (BAC) sequences representing approximately 80% of the genome had been analyzed, almost 60% of the genome was found to contain duplicated genes and regions (Blanc *et al.*, 2000). This phenomenon could only be explained by a complete genome duplication event, an opinion shared by the *Arabidopsis* Genome Initiative (2000).

Comparative studies of BACs between *Arabidopsis* and soybean (Grant *et al.*, 2000), and between *Arabidopsis* and tomato (Ku *et al.*, 2000), led to similar conclusions. In the latter case, two complete genome duplications were proposed: one 112 million years ago and another 180 million years ago. After dating duplicated blocks through a molecular clock analysis, Vision *et al.* (2000) also rejected the single-genome duplication hypothesis put forward by the *Arabidopsis* Genome

Initiative (2000). Several different age classes among the duplicated blocks were found, ranging from 50 to 220 million years, and at least four rounds of large-scale duplications were postulated. One of these classes, dated to approximately 100 million years ago, grouped nearly 50% of all the duplicated blocks, suggesting a complete genome duplication at that time (Vision *et al.*, 2000). However, the dating methods used for these gene duplications were based on averaging evolutionary rates of different proteins, which was later criticized because of their high sensitivity to rate differences (Wolfe, 2001; Raes *et al.*, 2003). Nevertheless, Vision *et al.* (2000) had discovered multiplicons of greater than Level 2, which can only be explained by multiple duplication events. By applying the novel techniques described earlier to detect heavily degenerated block duplications in *Arabidopsis*, Simillion *et al.* (2002) showed that the genome of this species had been reshaped by not one, but three entire genome duplication events. Recently, this result has also been confirmed through the construction and dating of evolutionary trees using genes from *Arabidopsis* and other plants (see Bowers *et al.*, 2003).

In stark contrast to *Arabidopsis*, where initial sequencing of the genome quickly revealed numerous duplicated segments, no clear evidence for ancient genome duplications had been reported for rice (*Oryza sativa*) until very recently, even though a paleopolyploid origin had been suggested for this species on several occasions (e.g., Goff *et al.*, 2002; Levy and Feldman 2002). Because the rice genome has now been completely sequenced (Goff *et al.*, 2002; Yu *et al.*, 2002), it is possible to apply the same approaches used in *Arabidopsis*. Based on a BAC assembly covering more than 70% of the genome sequence of *O. sativa*, the ADHoRe algorithm was applied to detect block duplications at the gene level. In addition to the detection of a large number of duplicated segments by direct comparison of all rice genomic scaffolds, a comparative approach using the genome sequence of *Arabidopsis* also yielded a set of ghost duplications, reflecting heavily degenerated duplicated segments. Of the 43 large block duplications (i.e., those with more than five anchor points), 34% of the total number of genes in these segments are retained as duplicates. When taking into account the estimated time of duplication, this fraction of retained gene duplicates is very similar to what has been observed in *Arabidopsis* and yeast (28% and 25%, respectively; Wolfe and Shields 1997; Simillion *et al.*, 2002), which seems to indicate similar rates of gene loss after duplication events.

When examining all multiplicons present in the rice genome through nonhidden, hidden, and ghost duplications, it is apparent that approximately 1.3% of the genome resides in multiplicons higher than Level 2. This implies that, given the quality of the current rice genomic data, a very small number of chromosomal regions have been involved in multiple duplication events. Again, this is very different from the situation in *Arabidopsis*, where the majority of chromosomal regions have been involved in multiple duplication events (Vision *et al.*, 2000; Simillion *et al.*, 2002; Bowers *et al.*, 2003).

In order to answer the question of whether rice is an ancient polyploid, Vandepoele *et al.* (2003) compared the duplication history of *Arabidopsis* and rice by plotting the total number of gene pairs in both species against their genetic distance inferred from the nucleotide substitutions at silent sites. When all duplicated gene pairs in *Arabidopsis* and rice are plotted as a function of K_s, the shape and height of the two curves are quite different (Vandepoele *et al.*, 2003). In *Arabidopsis*, the number of duplicates with K_s values between 0.6 and 0.9 increases dramatically, which corresponds with a genome duplication about 40 to 75 million years ago, as previously reported (Lynch and Conery, 2000; Simillion *et al.*, 2002; Blanc *et al.*, 2003; Bowers *et al.*, 2003). A small but significant increase can also be observed for rice duplicates with K_s values between 0.6 and 1.1. Because the relative increase in the number of duplicates is much smaller in rice than in *Arabidopsis*, a complete genome duplication in rice was considered highly unlikely, with aneuploidy given as the preferred explanation. However, recent analysis of a better assembly of the rice genome does seem to provide evidence for the occurrence of a whole genome duplication in rice about 70 million years ago (Guyot and Keller, 2004; Paterson *et al.*, 2004).

LARGE-SCALE DUPLICATIONS IN THE EVOLUTIONARY PROCESS

The Maintenance of Duplicated Genes

Before considering the role of large-scale duplications in influencing patterns of evolution, it is important to briefly review what happens to the genes themselves after duplication. Specifically, it is useful to consider why duplicated genes might be preserved in the genome over long evolutionary time periods. Some of these concepts were covered in more detail in Chapter 5 with reference to smaller duplications, but they also apply to genes duplicated *en masse*. The possibilities described here include "neofunctionalization," "subfunctionalization," and functional shift owing to positive selection.

After duplication, the two copies of the gene are redundant, meaning that they perform the same function and that inactivation of one gene should have little or no effect on the biological phenotype (Nowak *et al.*, 1997; Gibson and Spring, 1998; Lynch and Conery, 2000; Gu *et al.*, 2003). Therefore, because one of the copies is freed from functional constraint, mutations in this gene will be selectively neutral and will most often turn the gene into a nonfunctional pseudogene. As discussed in Chapter 5, the hypothesis presented by Ohno (1970) and several of his predecessors that gen(om)e duplications are vital for evolutionary diversification was often based on the notion of "neofunctionalization." That is, instead of being rendered inactive, on rare occasions one of the copies may be converted to a novel gene with a new function by a fortuitous series of nondeleterious mutations (Ohno, 1970, 1973). Although this model has been widely adopted to explain the evolution of functionally novel genes, little evidence has actually been found to support this mechanism. Moreover, under Ohno's model, one might consider it unlikely that anciently duplicated genes still perform completely redundant functions, yet redundancy has been shown to be widespread in the genomes of complex organisms (Nowak *et al.*, 1997, and references therein; Gibson and Spring, 1998; Li *et al.*, 2003).

The more recent alternative "duplication-degeneration-complementation" (DDC) model provides some explanation as to why duplicate genes might be retained (Force *et al.*, 1999; Lynch and Force, 2000). As noted in the previous chapter, this model starts from the assumption that a gene can perform several different functions; for instance, genes are expressed in different tissues and at different times during development, which may be controlled by different DNA regulatory elements. When duplicated genes lose different regulatory subfunctions, each affecting different spatial and/or temporal expression patterns, they must complement each other by jointly retaining the full set of subfunctions that were present in the single ancestral gene. Therefore, degenerative mutations facilitate the retention of duplicate functional genes, where both duplicates now perform different but necessary subfunctions. Therefore, the DDC model predicts that the sum of the retained duplicates is equal to the total number of subfunctions performed by the ancestral gene.

In short, according to the DDC model of Force *et al.* (1999), degenerative mutations preserve rather than destroy duplicated genes, but also change, or at least restrict, their functions to make them more specialized. Such a mechanism may prove to apply to the retention of many different gene duplicates, and indeed an increasing number of genes expected to have been subfunctionalized is being described (e.g., Prince and Pickett, 2002; Van de Peer *et al.*, 2003).

It is not only genes expressed in different tissues or at different times that can be subfunctionalized. For example, Gibson and Spring (1998) argued that selection can prevent the loss of redundant genes (i.e., duplicates) if these genes code for components of multidomain/multimer proteins. This is because inferior copies of these genes (or rather, their gene products) might inhibit the proper working of the "original" gene product. This hypothesis might explain why many transcription factors (TFs), which often form dimers, in gene families of plants contain so many members, many of which are probably redundant (De Bodt *et al.*, 2003; J. Spring, personal communication).

Positive Darwinian selection can also be responsible for functional divergence between duplicated genes (e.g., Zhang *et al.*, 1998; Duda and Palumbi, 1999; Hughes *et al.*, 2000). Most studies that look for evidence of positive Darwinian selection[1]

[1]For a review of the computational methods used to detect positive selection in duplicated genes, see Raes and Van de Peer (2003).

compare the ratio of nonsynonymous (p_N) and synonymous (p_S) substitutions (Hughes, 1999a; Nei and Kumar, 2000). In most genes, synonymous substitutions occur at a higher rate than nonsynonymous ones, because purifying selection prevents amino acid sequence changes (which are mostly disadvantageous). Under neutral evolution, the rates of synonymous and nonsynonymous substitutions are expected to be equal (Kimura, 1983). However, under positive Darwinian selection, amino acid replacements (i.e., nonsynonymous mutations) are favored. As a result, nonsynonymous mutations occur at a faster rate than synonymous mutations, as has been shown previously for genes and proteins such as primate lysozymes (Messier and Stewart, 1997), pregnancy-associated glycoproteins (Hughes *et al.*, 2000), primate ribonuclease genes (Zhang and Nei, 2000), conotoxins (Duda and Palumbi, 1999), opsins (Yokoyama *et al.*, 2000; Terai *et al.*, 2002), MYB DNA binding proteins (Jia *et al.*, 2003), and many others (see Endo *et al.*, 1996, and references therein).

Overall, the number of examples of the evolution of new and potentially adaptive functions in duplicated genes is, although growing, still quite small. Some of the more notable examples are the antifreeze proteins in Antarctic fishes (Cheng and Chen, 1999), color vision in new-world monkeys (Dulai *et al.*, 1999), thermal adaptation in *Escherichia coli* (Riehle *et al.*, 2001), and RNA digestion in colobine monkeys (Zhang *et al.*, 2002) (see also Chapter 5). Of course, large-scale gene or complete genome duplications, by whatever means, would provide an enormous number of "extra" genes with the potential to evolve new functions.

Which Genes Are Maintained, and Why?

The recent analyses of complete genome sequences have indicated that large-scale gene duplication has probably been rampant during the evolution of plants, fungi, and animals. It is tempting to speculate on the importance of such events for the biological evolution of these organisms. Indeed, as discussed earlier, it is to be expected that such major duplication events have been responsible for important evolutionary transitions and/or adaptive radiations of species (see also Chapters 5 and 11). However, providing hard evidence for direct correlations between large-scale gene duplication events and major leaps in evolution is not straightforward. An important first step in demonstrating that gene duplication events have indeed been of major importance for biological evolution would be to show which (kinds of) genes have generally been retained after gene duplication events.

For bacteria, this is relatively easy to do. With many complete genomes at hand, as well as more reliable genome annotations, it is possible to study which functional classes of genes show an excess of retained genes after duplication. Recent analysis of the functional classification of duplicated genes in bacteria, mainly created by small-scale duplication events such as tandem and operon duplications, revealed a preferential enrichment in functional classes that are involved in transcription, metabolism, and defense mechanisms (Gevers *et al.*, 2004).

Based on such analyses, it is also possible to consider links between gene retention and specific observations regarding the evolution and adaptation of organisms. For example, in the paranome of mycobacteria, two functional classes with an excess of retained duplicated genes are prominent, namely "lipid transport and metabolism" and "secondary metabolites biosynthesis, transport and catabolism" (Gevers *et al.*, 2004). Regarding the fatty acid metabolism, this is in agreement with the complex nature of the *Mycobacterium* cell wall and might reflect adaptive evolution of the bacterial cell surface. The case of *Borrelia burgdorferi* (the Lyme disease spirochete) is also informative in this context. In this species, the biased retention of duplicated motility genes and chemotaxis genes, together comprising more than 6% of its proteome, also appears to be biologically significant. Because *B. burgdorferi* lacks recognizable virulence factors, its ability to migrate to distant sites in its tick and mammalian hosts is probably dependent on a robust chemotaxis response (Fraser *et al.*, 1997). It has been suggested that multiple chemotaxis genes can be differentially expressed under varied physiological conditions or that different flagellar systems exist, requiring different chemotaxis systems (Fraser *et al.*, 1997).

Unfortunately, such analyses are much less straightforward in eukaryote genomes, in particular when the goal is to link gene retention with large-scale gene or entire genome duplication. For *Arabidopsis*, mathematical models are under development that will describe and simulate the retention of gene duplicates through time. Such models assume a constant "background" birth rate of new duplicates on which the three genome duplication events inferred for the *Arabidopsis* genome can be superimposed (Simillion *et al.*, 2002). Furthermore, this can allow different large-scale gene duplication events to have different decay rates with respect to each other and with respect to the continual background duplication process. Modeling both the continual mode of gene duplication as well as large-scale gene duplication events will also allow a comparison of the retention (and decay) of duplicates following large-scale duplication events for different functional categories of genes. It is hoped that this will provide a list of genes or gene categories that have been most important in driving evolution after duplication.

The Maintenance of Duplicated Genomes

If Ohno's proposition were true—that redundant genes, produced during large-scale gene duplication events, evolve previously nonexistent functions important for the evolution of phenotypic "complexity"—then traces of such events should be uncovered when the genomes of "complex" organisms are analyzed. Thanks to recent advances in genome sequencing and bioinformatics, it is

now recognized that many eukaryotes have undergone large-scale duplications of chromosomal segments and/or entire genomes (Wolfe and Shields, 1997; *Arabidopsis* Genome Initiative, 2000; Wolfe, 2001; Simillion *et al.*, 2002; Blanc *et al.*, 2003; Bowers *et al.*, 2003; Vandepoele *et al.*, 2003, 2004a). However, although duplicated genes and genomes may provide the raw material for evolutionary diversification, and functional divergence of duplicated genes (by several possible mechanisms) might offer a selective advantage to polyploids over a long time period, it is not yet clear how a partially or fully duplicated genome proves beneficial for an organism shortly after the duplication event. In other words, if a new genome doubling is to survive long enough to exert its long-term evolutionary effects, it must provide an immediate selective advantage that allows it to become established. There are several ways in which newly duplicated genomes might fulfill this requirement.

An important characteristic of duplicated genes is that they can buffer the genome against environmental perturbations and mutations, because when one copy of the gene is somehow inactivated, another with the same or similar function can be used instead. For example, Gu *et al.* (2003) have studied the effects of duplicated genes on the "fitness" of individuals of the budding yeast *Saccharomyces cerevisiae*. Based on functional data at the whole-genome level, the knocking out of single-copy genes was shown to generally reduce fitness more severely than deleting one gene of a pair of duplicates. As expected, duplicated genes that are highly similar in sequence are better at compensating for each other than duplicates whose sequences have diverged further. In conclusion, the study of Gu *et al.* (2003) demonstrates that duplicated genes may play an important role in genetic robustness against null mutations.

In plants, polyploidy often has immediate phenotypic effects with potential consequences for fitness, such as increased cell and organ size, faster growth, and increased capacity for invading new habitats (e.g., Osborn *et al.*, 2003) (see Chapter 7). In many cases, such differences in phenotype are probably caused by increased variation in dosage-regulated gene expression (Guo *et al.*, 1996). The fact that most ancient polyploids are thought to have been formed through allopolyploidy rather than autopolyploidy (Spring, 2003) is also relevant in this regard. Specifically, the combination of different genomes can lead to "hybrid vigor," placing the newly formed polyploid at a selective advantage compared to closely related diploid organisms.

Certainly, the prominence of polyploidy in flowering plants (see Chapter 7) implies that it has some adaptive significance, and hybridization has long been considered to be a significant evolutionary force that creates opportunities for adaptive evolution and speciation (Anderson, 1949; Ehrendorfer, 1980; Arnold, 1997; Ramsey and Schemske, 2002; Osborn *et al.*, 2003). Recently, Rieseberg *et al.* (2003) provided evidence that hybridization can play a key role in adaptation. These authors have employed several approaches to study the role of

hybridization in ecological adaptation and speciation in sunflowers, and showed that hybridization facilitated ecological divergence. In accordance with Spring (2003), they suggested that hybridization provides a mechanism for large and rapid adaptive transitions, made possible by the genetic variation at hundreds or thousands of genes in a single generation. Amores *et al.* (1998) and Wittbrodt *et al.* (1998) have also suggested that the potentially more complex genomic architecture of fishes resulting from an additional genome duplication might have permitted them to adapt and speciate more quickly in response to changing environments. Many studies have indeed shown that speciation can occur very rapidly in fishes, with the best known case being that of the African cichlids (Meyer *et al.*, 1990; Sturmbauer and Meyer, 1992; Meyer, 1993; Stiassny and Meyer, 1999; Wilson *et al.*, 2000).

In short, genome duplications may offer short-term selective advantages at each of the molecular, phenotypic, and ecological levels, in addition to influencing the long-term diversification of lineages.

Speciation and Divergent Resolution

Based on isozyme studies in ferns, Werth and Windham (1991) developed a model in which the "reciprocal silencing" of genes in geographically separated (allopatric) populations would promote speciation. Recently, this idea was revived in a model called "divergent resolution" (Lynch and Conery, 2000; Lynch and Force, 2000), in which the loss or silencing of gene duplicates may be even more important to the evolution of species diversity than the acquisition of new functions by the duplicated genes.

Divergent resolution occurs when different copies (on different chromosomes) of a duplicated gene are lost in allopatric populations, thereby creating genetic barriers to reproduction between them. Specifically, hybridization between such allopatric populations would produce an F_1 generation with one functional allele and one pseudogene at each of the duplicated loci (see Fig. 6.16). This in itself would not be problematic, but subsequent crosses between F_1 individuals would produce individuals with between zero and four alleles at the duplicated loci (Werth and Windham, 1991; Lynch and Force, 2000; Taylor *et al.*, 2001). Selection against F_2 individuals with more or fewer than two alleles per locus might provide a genetic environment in which speciation alleles (e.g., alleles for assortative mating) would be favored. Therefore, large-scale gene duplications might bring about rapid divergence because natural selection would favor speciation over hybridization in populations fixed for different copies of a duplicated locus.

Genome duplications produce an enormous number of gene duplicates that could be divergently resolved, with such genes potentially playing a prominent role in the generation of biodiversity by promoting the origin of postmating reproductive

barriers (Fig. 6.16). In this respect, it is noteworthy that in both ray-finned fishes and flowering plants there is a strong indication for a polyploidization event that seems to coincide with a massive diversification of novel lineages (Bowers *et al.*, 2003; Simillion *et al.*, 2003; Taylor *et al.*, 2003). On the other hand, whereas several studies have shown variation among populations in the retention of

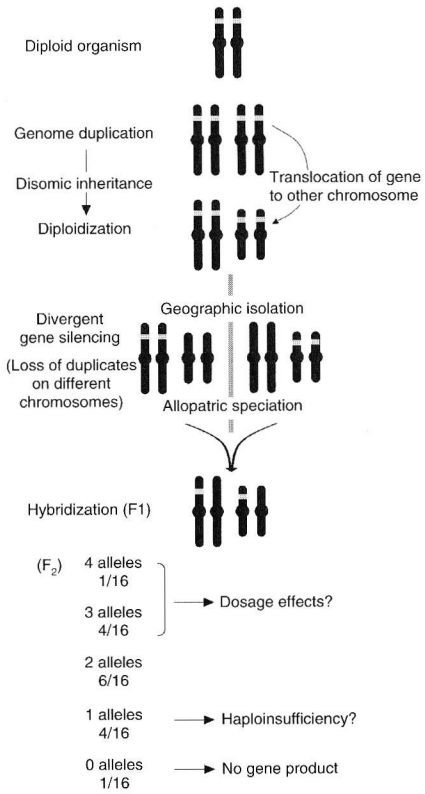

FIGURE 6.16 The role of duplications in speciation by divergent resolution. Gray bands represent a locus that is duplicated (along with all other loci) during a tetraploidization event. In this hypothetical example, diploidization is driven by a reciprocal translocation (to a different chromosome, depicted by a change in chromosome size). If individuals from isolated populations mate, their hybrid progeny would be heterozygous, possessing a functional allele (gray) and a pseudogene (black) at each locus of the duplicated gene. Crosses between the F_1 individuals produce some (approximately 6%) F_2 individuals with only pseudogenes at both of the loci in question, and therefore lacking viability and/or fertility. Others would receive from one allele, which might reduce functionality when the gene product from one functional allele is inadequate to support normal function (a phenomenon called "haploinsufficiency"), to three or four functional alleles, which might have a negative dosage effect. All these might lead to postmating reproductive isolation (Lynch and Force, 2000).

duplicated loci (reviewed by Taylor *et al.*, 2001), none has uncovered the pattern of gene loss predicted by the model. The tetraploid fish family Salmonidae (e.g., trout, salmon, char), which has many more species than its diploid sister group Esocidae (pike, pickerel, mudminnows), would be one good group in which to look for evidence of speciation owing to divergent resolution.

CONCLUDING REMARKS AND FUTURE PROSPECTS

It is becoming increasingly apparent that large-scale duplication events have featured prominently in many taxa, even those with small genomes. As the pace of complete genome sequencing continues to quicken, detailed investigations of this issue will become possible in an ever-widening array of species.

One important challenge for the next wave of studies will obviously be to identify additional duplication events themselves. This will be facilitated by the continued refinement of existing analytical techniques, as well as the development of new ones. It will also be important to discern the mechanisms responsible for these large-scale genomic events and to provide accurate estimates of the timings at which they have taken place. An understanding of the process of rediploidization, in which major duplication events are functionally undone, is also an area of considerable interest. The nature of the gene duplicates that persist, and the process by which duplicate pairs may diverge, is likewise a subject that is only beginning to be understood.

Perhaps most important of all will be the gaining of new insights, from the comparative study of many different genomes, regarding the evolutionary implications of large-scale gene and genome duplications. This involves considerations at several different levels, including impacts at the level of individual genes and entire genomes, the phenotypic and population-level consequences for the first organisms to exhibit the newly duplicated configuration, the possible input into the speciation process, and the long-term implications for major patterns of diversification.

This is indeed an exciting time in the study of genome biology, and one that will undoubtedly continue to alter the understanding of the evolutionary process at both genomic and geological scales.

REFERENCES

Abi-Rached L, Gilles A, Shiina T, *et al.* 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat Genet* 31: 100–105.

Altschul SF, Madden TL, Schaffer AA, *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.

Amores A, Force A, Yan YL, *et al.* 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282: 1711–1714.
Anderson E. 1949. *Introgressive Hybridization*. New York: Wiley.
Aparicio S, Chapman J, Stupka E, *et al.* 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297: 1301–1310.
Aparicio S, Hawker K, Cottage A, *et al.* 1997. Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. *Nat Genet* 16: 79–83.
Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.
Arnold ML. 1997. *Natural Hybridization and Evolution*. Oxford: Oxford University Press.
Bailey JA, Gu Z, Clark RA, *et al.* 2002. Recent segmental duplications in the human genome. *Science* 297: 1003–1007.
Benton MJ. 1990. Phylogeny of the major tetrapod groups: morphological data and divergence dates. *J Mol Evol* 30: 409–424.
Blanc G, Barakat A, Guyot R, *et al.* 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* 12, 1093–1101.
Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* 13: 137–144.
Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.
Brown TA. 1999. *Genomes*. Oxford: BIOS Scientific Publishers.
Calabrese PP, Chakravarty S, Vision TJ. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* 19: SI74–SI80.
Carroll RL. 1988. *Vertebrate Paleontology and Evolution*. New York: W. H. Freeman and Co.
Cheng CHC, Chen L. 1999. Evolution of an antifreeze glycoprotein. *Nature* 401: 443–444.
Conery JS, Lynch M. 2001. Nucleotide substitutions and the evolution of duplicate genes. *Pac Symp Biocomput*, 167–178.
De Bodt S, Raes J, Florquin K, *et al.* 2003. Structural annotation and evolutionary analysis of type I MADS box transcription factors in plants. *J Mol Evol* 56, 573–586.
Dietrich FS, Voegeli S, Brachat S, *et al.* 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304: 304–307.
Duda TF, Palumbi SR. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc Natl Acad Sci USA* 96: 6820–6823.
Dujon B, Sherman D, Fisher G. 2004. Genome evolution in yeasts. *Nature* 430: 35–44.
Dulai KS, von Dornum M, Mollon JD, Hunt DM. 1999. The evolution of trichromatic colour vision by opsin gene duplication in new world and old world primates. *Genome Res* 9: 629–638.
Durand D. 2003. Vertebrate evolution: doubling and shuffling with a full deck. *Trends Genet* 19: 2–5.
Easteal S, Collet C. 1994. Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. *Mol Biol Evol* 11: 643–647.
Ehrendorfer F. 1980. Polyploidy and distribution. In: Lewis WH ed. *Polyploidy—Biological Relevance*. New York: Plenum Press, pp. 45–60.
Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* 13: 685–690.
Escriva H, Manzon L, Youson J, Laudet V. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol* 9: 1440–1450.
Force A, Lynch M, Pickett FB, *et al.* 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545.
Fraser CM, Casjens S, Huang WM, *et al.* 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390: 580–586.
Fried C, Prohaska SJ, Stadler PF. 2003. Independent *Hox*-cluster duplications in lampreys. *J Exp Zool (Mol Dev Evol)* 299: 18–25.

Friedman R, Hughes AL. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol* 20: 154–161.
Furlong RF, Holland PWH. 2002. Were vertebrates octoploids? *Philos Trans R Soc Lond* B 357: 531–544.
Gates MA, Kim L, Cardozo T, *et al.* 1999. A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. *Genome Res* 9: 334–347.
Gaut BS. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res* 11: 55–66.
Gehring WJ. 1998. *Master Control Genes in Development and Evolution: The Homeobox Story*. New Haven, CT: Yale University Press.
Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* 12: 148–154.
Gibson TJ, Spring J. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet* 14: 46–49.
Goff SA, Ricke D, Lan TH, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92–100.
Grant D, Cregan P, Shoemaker RC. 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc Natl Acad Sci USA* 97: 4168–4173.
Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet* 31: 205–209.
Gu Z, Steinmetz LM, Gu X, *et al.* 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* 421: 63–66.
Guo M, Davis D, Birchler JA, *et al.* 1996. Dosage effect on gene expression in a maize ploidy series. *Genetics* 142: 1349–1355.
Guyot R, Keller B. 2004. Ancestral genome duplication in rice. *Genome* 47: 610–614.
Haldane JBS. 1933. The part played by recurrent mutation in evolution. *Am Nat* 67: 5–19.
Holland PW. 1997. Vertebrate evolution: something fishy about *Hox* genes. *Curr Biol* 7: R570–R572.
Holland PW, Garcia-Fernandez J. 1996. *Hox* genes and chordate evolution. *Dev Biol* 173: 382–395.
Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development* 120 (Suppl.): 125–133.
Hughes AL. 1999a. *Adaptive Evolution of Genes and Genomes*. Oxford: Oxford University Press.
Hughes AL. 1999b. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J Mol Evol* 48: 565–576.
Hughes AL, da Silva J, Friedman R. 2001. Ancient genome duplications did not structure the human *Hox*-bearing chromosomes. *Genome Res* 11: 771–780.
Hughes AL, Green JA, Garbayo JM, Roberts RM. 2000. Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proc Natl Acad Sci USA* 97: 3319–3323.
Irvine S, Carr JL, Bailey WJ, *et al.* 2002. Genomics analysis of *Hox* clusters in the sea lamprey *Petromyzon marinus*. *J Exp Zool* 249: 47–62.
Jia L, Clegg MT, Jiang T. 2003. Excess of non-synonymous substitutions suggest that positive selection episodes occurred during the evolution of DNA-binding domains in the *Arabidopsis* R2R3-MYB gene family. *Plant Mol Biol* 52: 627–642.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428: 617–624.
Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press.
Koszul R, Caburet S, Dujon B, Fischer G. 2004. Eukaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* 23: 234–243.
Ku HM, Vision T, Liu J, Tanksley SD. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci USA* 97, 9121–9126.
Larhammar D, Lundin LG, Hallbook F. 2002. The human *Hox*-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res* 12: 1910–1920.

Levy A, Feldman M. 2002. The impact of polyploidy on grass genome evolution. *Plant Physiol* 130: 1587–1593.
Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 36: 96–99.
Li WH, Gu Z, Cavalcanti ARO, Nekrutenko A. 2003. Detection of gene duplications and block duplication in eukaryotic genomes. *J Struct Funct Genomics* 3: 27–34.
Lin X, Kaul S, Rounsley S, *et al.* 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402: 761–768.
Lundin LG, Larhammar D, Hallbook F. 2003. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J Struct Funct Genomics* 3: 53–63.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473.
Málaga-Trillo E, Meyer A. 2001. Genome duplications and accelerated evolution of *Hox* genes and cluster architecture in teleost fishes. *Am Zool* 41: 676–686.
Martin A. 2001. Is tetralogy true? Lack of support for the "one-to-four rule." *Mol Biol Evol* 18: 89–93.
Mayer K, Schüller C, Wambutt R, *et al.* 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402: 769–777.
McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet* 31: 200–204.
Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysosymes. *Nature* 385: 151–154.
Meyer A. 1993. Phylogenetic relationships and evolutionary processes in East African cichlids. *Trends Ecol Evol* 8: 279–284.
Meyer A, Kocher TD, Basasibwaki P, Wilson A. 1990. Monophyletic origin of Lake Victoria Africa cichlid fishes suggested by mitochondrial DNA sequences. *Nature* 347: 550–663.
Meyer A, Schartl M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol* 11: 699–704.
Naruse K, Fukamachi S, Mitani H, *et al.* 2000. A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. *Genetics* 154: 1773–1784.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3: 418–426.
Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford: Oxford University Press.
Nowak MA, Boerlijst MC, Cooke J, Maynard Smith J. 1997. Evolution of genetic redundancy. *Nature* 388: 167–171.
Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer Verlag.
Ohno S. 1973. Ancient linkage groups and frozen accidents. *Nature* 244: 259–262.
Osborn TC, Pires JC, Birchler JA, *et al.* 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet* 19: 141–147.
Panopoulou G, Hennig S, Groth D, *et al.* 2003. New evidence for genome-wide duplications at the origin of vertebrates using an Amphioxus gene set and completed animal genomes. *Genome Res* 13: 1056–1066.
Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* 101: 9903–9908.
Paterson AH, Lan TH, Reischmann KP, *et al.* 1996. Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nat Genet* 14: 380–382.
Postlethwait JH, Woods IG, Ngo-Hazelett P, *et al.* 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res* 10: 1890–1902.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* 3: 827–837.
Raes J, Van de Peer Y. 2003. Gene duplications, the evolution of novel gene functions, and detecting functional divergence of duplicates *in silico*. *Appl Bioinformatics* 2: 92–101.
Raes J, Vandepoele K, Simillion C, *et al.* 2003. Investigating ancient duplication events in the *Arabidopsis* genome. *J Struct Func Genomics* 3: 117–129.
Ramsey J, Schemske DW. 2002. Neopolyploidy in flowering plants. *Annu Rev Ecol Syst* 33: 589–639.
Riehle MM, Bennette AF, Long AD. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc Natl Acad Sci USA* 98: 525–530.
Rieseberg LH, Raymond O, Rosenthal DM, *et al.* 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* 301: 1211–1216.
Robinson-Rechavi M, Marchand O, Escriva H, *et al.* 2001a. Euteleost fish genomes are characterized by expansion of gene families. *Genome Res* 11: 781–788.
Robinson-Rechavi M, Marchand O, Escriva H, Laudet V. 2001b. An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Curr Biol* 11: R458–R459.
Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* 12: 85–94.
Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Curr Opin Microbiol* 2: 548–554.
Shu DG, Luo HL, Conway Morris S, *et al.* 1999. Lower Cambrian vertebrates from south China. *Nature* 402: 42–46.
Simillion C, Vandepoele K, Saeys Y, Van de Peer Y. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res* 14: 1095–1106.
Simillion C, Vandepoele K, Van Montagu M, *et al.* 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 99: 13627–13632.
Skrabanek L, Wolfe KH. 1998. Eukaryote genome duplication: where's the evidence? *Curr Opin Genet Dev* 8: 694–700.
Spring J. 1997. Vertebrate evolution by interspecific hybridisations: are we polyploid? *FEBS Lett* 400: 2–8.
Spring J. 2003. Major transitions in evolution by genome fusions: from prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates. *J Struct Funct Genomics* 3: 19–25.
Stephens SG. 1951. Possible significance of duplication in evolution. *Adv Genet* 4: 247–265.
Stiassny MLJ, Meyer A. 1999. Cichlids of the African Rift Lakes. *Sci Am* February: 64–69.
Sturmbauer C, Meyer A. 1992. Genetic divergence, speciation and morphological stasis in a lineage of African cichlid fishes. *Nature* 358: 578–581.
Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol* 12: 823–833.
Taylor J, Braasch I, Frickey T, *et al.* 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res* 13: 382–390.
Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B* 356: 1661–1679.
Terai Y, Mayer WE, Klein J, *et al.* 2002. The effect of selection on a long wavelength-sensitive (LWS) opsin gene of Lake Victoria cichlid fishes. *Proc Natl Acad Sci USA* 99: 15501–15506.
Terryn N, Heijnen L, De Keyser A, *et al.* 1999. Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. *FEBS Lett* 445: 237–245.
Van de Peer Y, Taylor JS, Braasch I, Meyer A. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol* 53: 436–446.
Van de Peer Y, Taylor JS, Meyer A. 2003. Are all fishes ancient polyploids? *J Struct Funct Genomics* 3: 65–73.

Vandepoele K, De Vos W, Taylor JS, *et al.* 2004a. Major events in the genome evolution of vertebrates: paranome age and size differs considerably between fishes and land vertebrates. *Proc Natl Acad Sci USA* 101: 1638–1643.

Vandepoele K, Saeys Y, Simillion C, *et al.* 2002a. A new tool for the automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res* 12: 1792–1801.

Vandepoele K, Simillion C, Van de Peer Y. 2002b. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet* 18: 606–608.

Vandepoele K, Simillion C, Van de Peer Y. 2003. Evidence that rice, and other cereals, are ancient aneuploids. *Plant Cell* 15: 2192–2202.

Vandepoele K, Simillion C, Van de Peer Y. 2004b. The quest for genomic homology. *Curr Genomics* 5: 299–308.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114–2117.

Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploidy pteridophytes resulting from silencing of duplicate-gene expression. *Am Nat* 137: 515–526.

Wilson AB, Noack-Kunnmann K, Meyer A. 2000. Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: sexual selection versus ecological diversification. *Proc R Soc Lond B* 267: 2133–2141.

Wittbrodt J, Meyer A, Schartl M. 1998. More genes in fish? *BioEssays* 20: 511–512.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2: 333–341.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708–713.

Woods IG, Kelly PD, Chu F, *et al.* 2000. A comparative map of the zebrafish genome. *Genome Res* 10: 1903–1914.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556 (available at http://abacus.gene.ucl.ac.uk/software/paml.html).

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17: 32–43.

Yokoyama S, Blow NS, Radlwimmer FB. 2000. Molecular evolution of color vision of zebra finch. *Gene* 259: 17–24.

Yu J, Hu S, Wang J, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92.

Zhang J, Nei M. 2000. Positive selection in the evolution of mammalian interleukin-2 genes. *Mol Biol Evol* 17: 1413–1416.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* 95: 3708–3713.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet* 30: 411–415.

Zhu M, Yu X. 2002. A primitive fish close to the common ancestor of tetrapods and lungfish. *Nature* 418: 767–770.

Zhu M, Yu X, Janvier P. 1999. A primitive fossil fish sheds light on the origin of bony fishes. *Nature* 397: 607–610.