8 Genomic Expansion by Gene Duplication

Dating back to the pre-molecular era (Muller 1940; Haldane 1933; Ohno 1970; reviewed in Taylor and Raes 2004), substantial attention has been given to the idea that gene duplication is the major mechanism for the origin of new gene functions, and it is now firmly established that the refashioning of duplicate genes is a major contributor to the origin of adaptive evolutionary novelties (numerous examples are cataloged in Ganfornina and Sanchez 1999; Patthy 1999b; True and Carroll 2002; Hurley et al. 2005; Irish and Litt 2005; Nei and Rooney 2005). However, these attention-grabbing examples need not be representative of the fates of *average* gene duplicates, as there is a logical distinction between the processes responsible for the initial establishment of duplicate genes and their secondary modification by mutation and natural selection. Virtually all new genes must arise from accidental duplications of preexisting genes or parts thereof, which implies an initial state of a single copy in a single member of the population. Thus, understanding the processes that facilitate the expansion versus contraction of gene number requires an appreciation of the molecular processes that give rise to duplication events and of the population genetic forces that influence the dynamics of newly arisen genes.

If beneficial mutations resulting in new gene functions were the primary means for preserving duplicate genes, then because such mutations are rare, species with enormous population sizes would be expected to carry the largest numbers of genes. That this is not the case immediately suggests that other defining forces must be at work. A key point to be made below is that random genetic drift and degenerative mutations commonly play a central role in the growth of gene number in populations that are sufficiently small in size. Although the preservation of duplicate genes by degenerative mutations may seem counterintuitive, we will see that this process is a nat-

ural outcome of the structure of eukaryotic genes. Once this principle is understood, the conclusion that phylogenetic changes in gene number may be substantially influenced by nonadaptive processes becomes inescapable, although as noted above, this need not imply that gene duplication is a minor player in the origin of new gene functions.

Prior to addressing the population genetic mechanisms by which duplicate genes evolve, the context of the problem will first be established by considering the rates at which such genes arise and the time periods over which they typically survive. Although substantial evidence suggests that many key evolutionary lineages of multicellular eukaryotes have experienced one or more complete genome doublings (polyploidization) at some time in the distant past (Wolfe 2001), it will be seen that gene duplication by smaller-scale processes is an ongoing feature of all organisms. The genome is a dynamic playing field on which new genes are continually arising via duplication events, with most being eliminated fairly quickly by drift and/or natural selection, some simply replacing their parental copies, and a few experiencing functional changes that ensure their long-term preservation along with their ancestral family members. Even in the absence of any net growth in genome size, this continual turnover of genes has further evolutionary implications, as it passively promotes the origin of microchromosomal rearrangements. In this sense, the gene duplication process provides fuel for both of the major engines of evolution: adaptive phenotypic change within lineages and the creation of new lineages by speciation.

The Evolutionary Demography of Duplicate Genes

The power of gene duplication as an evolutionary force depends on the rate at which duplicate genes arise. Although there is currently no simple way to estimate this rate directly in laboratory experiments, information from complete genome sequences provides an indirect approach (Lynch and Conery 2000, 2003a). Through comparative sequence analysis, the total pool of duplicate genes within a genome can be identified, and under the assumption that silent sites accumulate nucleotide changes at a relatively constant rate, the relative age of each duplicate pair can be estimated from the silent-site divergence between pair members. The resultant age distribution of duplicate pairs can then be used to estimate the average rates of origin and elimination of duplicate genes, using the same principles that demographers use to estimate birth and death rates of individuals in natural populations.

If it can be assumed that birth and death rates of duplicate genes have remained roughly constant over the time scale of observation, a particularly powerful analysis becomes possible. Under a steady-state birth/death process, the expected frequency of duplicates declines exponentially with age, with the time-zero intercept of the age distribution providing information on the birth rate and the slope providing information on the death rate. To see this, recall the simple model introduced in Chapter 3,

$$n_t = n_{t-1} + B(1 + n_{t-1}) - Dn_{t-1}$$

where n_t denotes the number of genomic copies of a gene at generation t (in excess of the baseline single-copy requirement for viability), B denotes the rate of gene birth (applied to the baseline and extra copies), and D denotes the rate of gene death (applied only to the excess copies, and otherwise assumed to be independent of copy number). Both B and D are stochastic variables, but averaging over a large pool of genes (the entire genome), some specific patterns can be predicted. At equilibrium ($n_t = n_{t-1}$), the expected number of excess copies per gene is $n_{tot} = B/(D - B)$, which is just a function of the ratio B/D. If $D >> B$ (justified by data presented below), $n_{tot} \approx B/D$, which is much less than 1, so the total birth rate per gene family, $B(1 + n_{tot})$, is close to B each generation. Furthermore, because a constant fraction D is lost each generation, the steady-state age distribution is close to

$$n_i = B(1 - D)^i$$

as graphed in Figure 8.1 (Lynch and Conery 2003a). Under this model, a log-linear plot of n_i versus i is expected to yield a straight line with the expected form

$$\log(n_i) = \log B + i \cdot \log(1 - D)$$

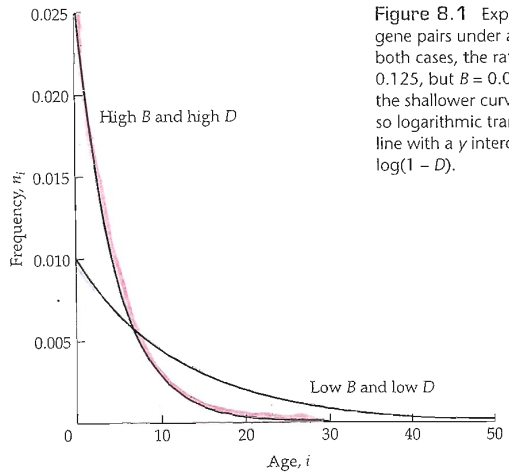

Figure 8.1 Expected age distributions of duplicate-gene pairs under a steady-state birth/death process. In both cases, the ratio of birth to death (B/D) is equal to 0.125, but $B = 0.025$ for the upper curve and $B = 0.01$ for the shallower curve. The plotted function is $n_i = B(1 - D)^i$, so logarithmic transformation of the y axis yields a straight line with a y intercept equal to log B and a slope equal to log$(1 - D)$.

Thus, by fitting a linear regression to the logarithmic transformation of the age distribution, estimates of the birth and death rates of duplicate genes can be acquired by setting the intercept and slope equal to log B and log(1 − D), respectively. The age distribution (n_i) for such an analysis can be acquired by querying an entire genome for the number of gene pairs of each age i (binning in units of silent-site divergence) and dividing by the total number of genes (not including the duplicates themselves).

We have already encountered a duplicate-gene age distribution for the human genome (see Figure 3.1) that closely approximates the exponential form suggested by the preceding model, and another such distribution for the *Caenorhabditis elegans* genome appears in Figure 8.2A. Many additional eukaryotic genomes exhibit this pattern, at least as a first approximation

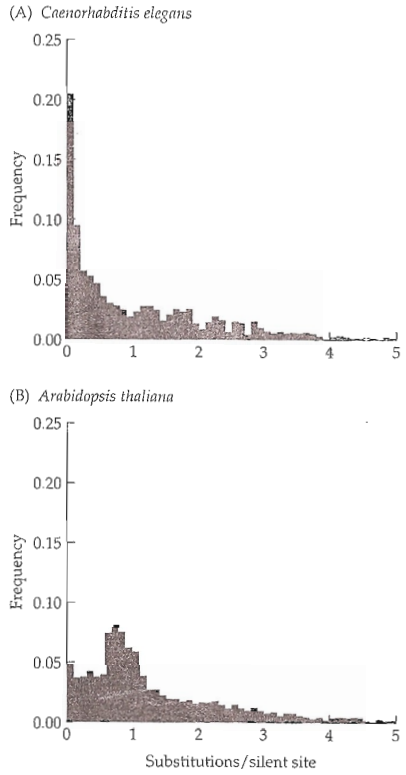

Figure 8.2 Age distributions of duplicate genes in the nematode *Caenorhabditis elegans* (A) and the plant *Arabidopsis thaliana* (B). The approximate sample sizes are 1100 and 3200, respectively. The large internal peak in the *Arabidopsis* age distribution is a reflection of an ancient polyploidization event. (From Lynch and Conery 2003a.)

(Lynch and Conery 2000, 2003a; Achaz et al. 2001; Blanc and Wolfe 2004). However, nonequilibrium situations, such as a "baby boom" resulting from a past polyploidization event, are sometimes inferred from an intermediate peak in the age distribution (Gu et al. 2002a; McLysaght et al. 2002; Jaillon et al. 2004; Vandepoele et al. 2004; Maere et al. 2005; Figure 8.2B). In principle, bulged distributions can also result from periods of reduced duplicate-gene loss.

Application of the steady-state solution to the genomes of diverse eukaryotic species yields birth rate estimates for duplicate genes that are generally in the range of 0.001 to 0.01 per gene on the time scale of 1% divergence (Table 8.1). There is no general phylogenetic pattern to these values, as both unicellular species and animals have average values of about 0.004, and the one land plant for which data are available, *Arabidopsis thaliana*, has an estimate of 0.003. It should be emphasized that these estimated rates of gene duplication apply to single genes in single individuals. They are not population-level rates of origin of new duplicates (which are equivalent to $2NB$, where N is the number of individuals in a diploid population), nor are they equivalent to fixation rates of new duplicates (which would be diminished by processes of duplicate-gene loss).

TABLE 8.1 Estimated rates of origin (B) and loss (D) of duplicate genes for eukaryotes

SPECIES	B	D	B/D
Unicellular species			
Plasmodium falciparum	0.0003	0.167	0.0018
Saccharomyces cerevisiae	0.0025	0.324	0.0077
Schizosaccharomyces pombe	0.0016	0.386	0.0042
Encephalitozoon cuniculi	0.0117	0.487	0.0240
Animals			
Homo sapiens	0.0049	0.081	0.0605
Mus musculus	0.0030	0.134	0.0224
Fugu rubripes	0.0043	0.189	0.0228
Caenorhabditis elegans	0.0028	0.136	0.0206
Drosophila melanogaster	0.0011	0.229	0.0048
Anopheles gambiae	0.0062	0.190	0.0326
Plants			
Arabidopsis thaliana	0.0032	0.033	0.0970

Source: Lynch and Conery 2003a.

Note: Both B and D are defined on a time scale of 1% divergence for silent sites. As noted in the text, the ratio B/D provides an estimate of the average number of excess copies per gene resulting from the stochastic birth/death process.

It is difficult to apply the age distribution approach to prokaryotic species because of their greatly reduced number of genes. However, a downwardly biased average estimate of B for 73 species of prokaryotes, which simply counts the number of duplicates with silent-site divergence of less than 1% and does not account for early gene loss, is about 0.002 (Lynch and Conery 2003a), well within the range of eukaryotic species. As many other indirect observations support the idea that rates of gene duplication (some involving horizontal transfer) are high in prokaryotes (Anderson and Roth 1977; Brenner et al. 1995; Lawrence and Ochman 1998; Bergthorsson and Ochman 1999; Lerat et al. 2005), the relatively small number of genes within prokaryotic genomes is a consequence of a relatively high rate of attrition of gene duplicates.

As a first approximation, these observations suggest that the rate of duplication of entire genes is only slightly less than the rate at which nucleotide substitutions occur at silent sites. Given this scaling and the fact that the amount of silent-site substitution per generation increases with generation time (see Chapter 4), it follows that the per-generation rate of gene duplication increases from unicellular to multicellular species. Recalling the estimates for silent-site divergence given in Figure 4.5, 1% silent-site divergence is roughly equivalent to 6×10^6, 8×10^5, and 2×10^5 generations in unicellular eukaryotes, invertebrates, and vertebrates, respectively, which in turn imply birth rate estimates of 0.06%, 0.5%, and 2% per gene per 10^6 generations for these three groups. Using an average rate of silent-site substitution for vascular plants derived from six independent studies (Gaut et al. 1996; Li 1997; Lynch 1997; Koch et al. 2000), 8.1 per site per billion years, and assuming two generations per year, the birth rate of gene duplicates in *Arabidopsis* is about 0.8% per gene per 10^6 generations.

There are a number of significant caveats with respect to these estimates. For example, most of them assume an equilibrium age distribution of duplicate genes and an approximately constant rate of silent-site divergence. Because duplicate genes generally have identical sequences at the time of origin and are also often in close spatial proximity, they are expected to be subject to stochastic homogenization by gene conversion. In principle, gene conversion between duplicate pairs could also be mediated by recombination with reverse-transcribed mRNAs, in which case highly expressed genes would be especially prone to rejuvenation (Pyne et al. 2005). Such events would result in a nonlinear relationship between the age of a pair and the rate of silent-site divergence until the level of divergence exceeded the point beyond which homology-dependent conversion is possible, potentially leading to overestimates of B.

Despite these concerns, the results reported above appear to be quite robust. Using rather different approaches, Cotton and Page (2005) obtained a human duplication rate estimate identical to that given above when a 20-year generation time was assumed. Furthermore, studies of segmental duplications in the human population, which reveal hundreds of presence/absence polymorphisms of duplication spans of up to several hundred kilobases in length (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005; Khaja et al. 2006; Redon et al. 2006), imply minimum duplication rates of about 2% per gene per million years (van Ommen 2005). In addition, Maere et al. (2005) arrived at a birth rate estimate for *Arabidopsis* identical to that given above for land plants. An exceedingly tiny estimate (~0.001%/gene/MY) reported for *Saccharomyces cerevisiae* (Gao and Innan 2004) is clearly an error, as the authors actually considered the rate of gene preservation (which is diminished by mechanisms of gene loss) rather than birth. Finally, because specific genes in *Drosophila* have been found to duplicate at rates as high as 1.0 to 100 per gene per 10^6 generations (Gelbart and Chovnick 1979; Shapira and Finnerty 1986), the indirect estimates suggested above are not biologically unrealistic and may be quite conservative.

Substantial differences in the loss rates of duplicate genes also exist between unicellular and multicellular species. On the time scale of 1% silent-site divergence, D averages 0.34 (SE = 0.07) for unicellular species, 0.18 (0.03) for invertebrates, 0.13 (0.03) for vertebrates, and 0.033 for *Arabidopsis*, but because 1% silent-site divergence requires many more generations in unicellular than in multicellular species, the per-generation loss rates are actually higher in multicellular species. On a time scale of 10^6 generations, estimated loss rates are ~0.04 for unicellular eukaryotes, 0.10 for invertebrates, and 0.28 for vertebrates and *Arabidopsis*, implying half-lives for duplicate genes in these three groups of 16.7, 7.0, and 2.5 millions of generations, respectively.

These results suggest that with respect to gene content, the eukaryotic genome is highly dynamic. As a consequence of a stochastic balance between gene birth and death rates, total genome size may remain approximately constant for long periods, but throughout such periods there is likely to be continual turnover with respect to the specific genes that are present in redundant copies. Given that the probability of a gene duplicating by the time a silent site experiences 0.01 substitutions is roughly 0.004, the rate of duplication per gene is about 40% of the rate of mutation per nucleotide site, raising the possibility that changes in gene content may often rival changes in gene sequence as a mechanism of phenotypic evolution. Indeed, on a time scale of roughly 250 million years, nearly every gene in a typical eukaryotic genome can be expected to duplicate at least once. On the other hand, as is the case for most amino acid replacement substitutions within genes, most duplicate genes appear to be evolutionarily short-lived, with typical half-lives of just a few million years.

Origins of segmental duplications

The molecular mechanisms by which duplicate genes arise are diverse, ranging from complete genome duplication (discussed in the following section) to more restricted duplications of smaller chromosomal regions. The latter,

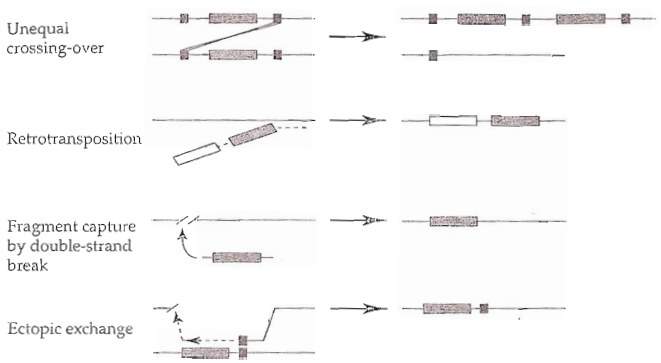

Figure 8.3 Four mechanisms for the origin of gene-sized duplications. Unequal crossing-over: A crossover occurs between two regions of sequence similarity (red) at nonhomologous sites, yielding one chromosome with a duplication and another with a deletion. Retrotransposition: A gene that has been transcribed along with an upstream retrotransposon (yellow) is inserted into a new site after reverse transcription of the mRNA intermediate (dashed line). Capture by double-strand break: An exogenous fragment containing a gene sequence (possibly originating by reverse transcription) is inserted into a chromosomal break point. Ectopic exchange: A double-strand break initiates a recombination event by invading a nonhomologous site, at which a gene copy is generated by strand extension prior to reannealing of the broken chromosome.

broadly defined as segmental duplications, arise by multiple pathways (Figure 8.3):

- Many newly arisen gene duplicates are tandemly associated with their parental copy, having arisen from local chromosomal events such as replication slippage or nonhomologous unequal crossing-over. Such local duplications can then be subsequently dispersed by various mechanisms of chromosomal rearrangement.
- Sloppy transcription of non-LTR retrotransposons occasionally leads to the replication of downstream genes and their reinsertion elsewhere in the genome, as emphasized in Chapter 3. The products of such events will often be nonfunctional (e.g., when the duplication span fails to incorporate key regulatory or coding domains), but the possibility also exists for the acquisition of entirely new expression patterns (e.g., when an insertion fortuitously incorporates regulatory sequences at the new site).
- Duplicates can originate via the capture of DNA inserts during the repair of double-strand breaks. Fragments of mitochondrial DNA and retrotransposon-derived cDNAs appear to be particularly common substrates

for this process (Ricchetti et al. 1999; Yu and Gabriel 1999; Lin and Waldman 2001a,b).

- Protruding ends of double-strand breaks may invade ectopic sites with short regions of homology, transiently using the invaded chromosome as a template for strand extension and then reattaching the two free ends (Gorbunova and Levy 1997). As in the case of LTR-derived duplications, whether insertions generated during double-strand break repair will contain one or more functional genes or simply be "dead on arrival" is entirely a matter of chance.

Structural analyses of the youngest cohorts of duplicate genes (<10% sequence divergence at silent sites) in the *C. elegans* genome provide some insight into these issues (Katju and Lynch 2003, 2006; Thomas 2006). In this species, the distribution of duplication-span lengths is highly L-shaped, with a mean of just 1.4 kb and only 30% of duplication events exceeding 2.5 kb in length (Figure 8.4). Thus, because the average length of a *C. elegans* gene from start codon to termination codon is about 1.9 kb, only 50% of newborn duplicates appear to be complete. Approximately 20% of newborn *C. elegans* genes are partial, in the sense that one member of the pair (presumably the parental copy) contains unique exonic DNA, and 30% are chimeric, with each member of the pair containing unique exons. The degree to which partial or chimeric duplicates are operable remains to be determined, but

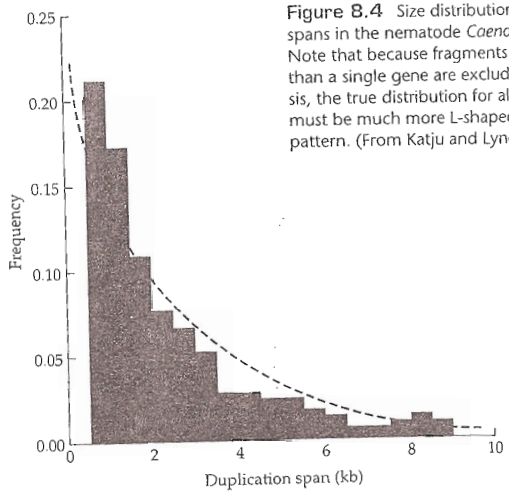

Figure 8.4 Size distribution of duplication spans in the nematode *Caenorhabditis elegans*. Note that because fragments containing less than a single gene are excluded from this analysis, the true distribution for all duplication spans must be much more L-shaped than the observed pattern. (From Katju and Lynch 2003.)

because the relative frequencies of such genes do not radically change in older cohorts, and because the expression of such genes is just as frequent as that of complete duplicates, some functional role seems likely.

Almost all (~90%) newborn duplicates in *C. elegans* are tandemly arranged. However, although tandem duplications are generally thought to arise via unequal crossing-over, which should lead to tail-to-head orientations of adjacent members, about 70% of newly arisen *C. elegans* duplicates are inverted and on opposite strands in tail-to-tail or head-to-head orientations. Tandem inversions can arise during replication if the DNA polymerase transiently switches strands, moving in the opposite direction on the complementary strand for a while, before switching back to the original strand (Bi and Liu 1996; Lin et al. 2001). Such inverted duplicates may be more stable than direct tandem repeats, which may be easily lost through the very same mechanisms that lead to their creation. L-shaped length distributions and inverted duplications are commonly observed in other species (Fischer et al. 2001; Bensasson et al. 2003; Thomas et al. 2004; Zhang et al. 2005).

Whole genome duplication

Although the previous analyses have focused on segmental duplications involving one to a few genes, entire genomes are sometimes duplicated via polyploidization events. Over the course of evolutionary history, whole-genome duplication events have been fairly common in plants, and although they are much rarer in animals (Ramsey and Schemske 1998; Otto and Whitton 2000), they nevertheless appear to occupy key positions in the animal phylogeny (Figure 8.5). Remarkably, as discussed below, the lineages of most of the model systems adopted by molecular, cell, and developmental biologists may have experienced at least one polyploidization event in the distant past.

The mechanisms of polyploidization are varied. Autopolyploids originate endogenously, with all alleles at a locus deriving from the same species, whereas allopolyploids arise via hybridization events. Both mechanisms often involve the participation of unreduced gametes, either at the outset in the case of autopolyploidy or subsequent to hybridization in the case of allopolyploidy. The initial stages of such events are made difficult by the fact that the incipient polyploid individual will generally be embedded within a population of diploids, and hence confronted with a high likelihood of yielding progeny with intermediate ploidy levels, which can inhibit the production of viable gametes. For example, the mating of a tetraploid with a diploid will produce triploid offspring that experience major problems during meiosis. Self-fertilization provides a simple mechanism for bypassing such problems in land plants, but most animals are obligate outcrossers.

Documenting an ancient polyploidization event can be difficult for several reasons. A clear signature of polyploidy is the presence of long colinear duplication spans of genes or, ideally, entire chromosomes. However, as

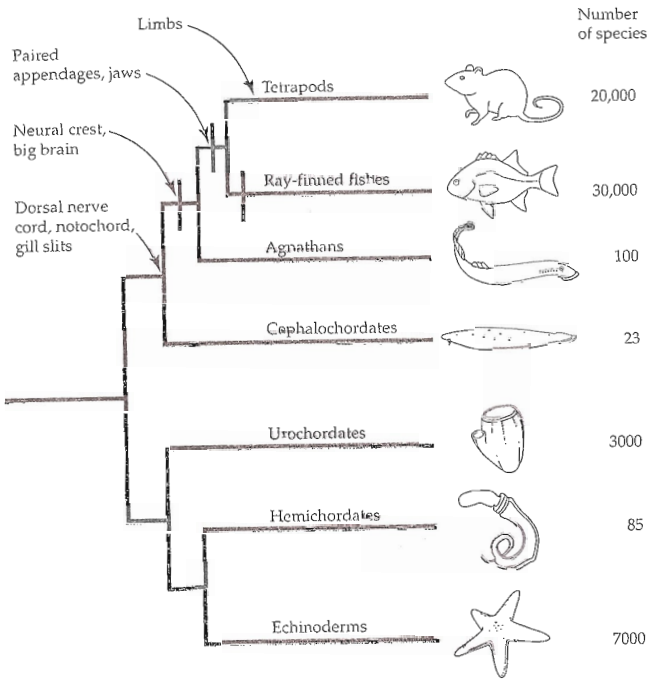

Figure 8.5 The chordate phylogeny. Lineage-specific morphological innovations are shown to the left, and numbers of species in each lineage are shown to the right. Blue bars denote periods of hypothesized genome duplication. The exact positions of the two duplications preceding the evolution of jawed vertebrates are uncertain, as is their polyploid origin; the duplication at the right denotes a polyploidization event deep in the ray-finned fish lineage, which has been followed by secondary polyploidization events in various fish lineages (not shown).

described above, many duplicate genes are eventually lost, leaving large gaps in what would otherwise be continuous stretches of genes. Secondary rearrangements of chromosomal segments (e.g., inversions and translocations) can further obscure the signature of polyploidization, as can background segmental duplications that increase the number of copies of genes beyond the expectation based on polyploidy alone. Bioinformaticians have attempted to grapple with these problems by factoring out apparently young gene duplicates and simultaneously analyzing several species, but acquir-

ing strong evidence for ancient polyploidization events is still a formidable task. Here, we briefly highlight four putative eukaryotic genome duplication events, all involving lineages that are central to experimental biology.

Although polyploidy is often regarded as a much more common feature of multicellular than unicellular species, the first detailed genomic analysis of a polyploidization event derived from the discovery of a large number of duplicate chromosomal fragments in the yeast *S. cerevisiae* (Wolfe and Shields 1997). Only about 8% of the originally duplicated genes survive today, and there has been substantial secondary movement of chromosomal segments (Figure 8.6). Nevertheless, the fact that nearly all gene pairs within surviving fragments have the same orientation with respect to centromeres supports the idea that these segments originally arose via whole-genome duplication, a conclusion that is also consistent with phylogenetic analysis (Langkjaer et al. 2003). More recent work has revealed an even more striking case of polyploidization in the unicellular ciliate (*Paramecium tetraurelia*), where an apparent three rounds of whole-genome duplication has yielded ~40,000 genes (Aury et al. 2006).

The *Arabidopsis* genome project was initiated on the assumption that this species would have a simple genome relative to that of other plants, but it soon became apparent that the chromosomal contents of this species reflect a complex history of duplication events. The bulge in the age distribution of *Arabidopsis* gene duplicates appearing at an average level of silent-site divergence of 0.8 (see Figure 8.2) suggests an ancient phase of genome amplification. Because a very large fraction of duplicate genes of this age will have been lost, and because the silent sites of the remaining pairs are saturated with mutations, it is difficult to infer the mechanisms underlying such ancient bouts of gene amplification from age distributions alone. However, additional work that takes chromosomal positions into account suggests that the *Arabidopsis* genome is a product of at least two, and probably three, polyploidization events (Simillion et al. 2002; Bowers et al. 2003): the first prior to the monocot–dicot split; the second between 160 and 230 MYA (after the divergence of monocots and dicots); and the third between 20 and 85 MYA. An earlier suggestion of a fourth round of polyploidization (Vision et al. 2000) has not been upheld. Additional observations on other angiosperm groups suggest the occurrence of numerous independent polyploidization events in descendent sublineages (Wendel 2000; Paterson et al. 2004). Large-scale partial genome duplications have also occurred in plants. In rice, for example, a single-chromosome duplication appears to have occurred around 54 MYA, the approximate time of origin of the major grass lineages (Vandepoele et al. 2003).

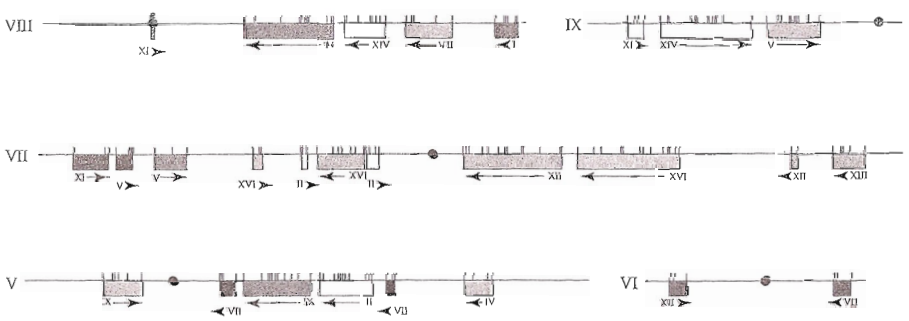

Figure 8.6 Examples of remnant chromosomal segments derived from an ancient whole-genome duplication in the yeast *S. cerevisiae*. Roman numerals denote chromosome numbers. Colored blocks of genes denote groups of genes (indicated by tick marks) for which duplicated spans exist on another chromosome (given by the Roman numeral underlying the block). Arrows denote the relative orientations of the full set of genes within a block. The solid circles demarcate centromeres. The mixture of blocks of different ancestries on individual chromosomes presumably results from chromosomal rearrangement subsequent to polyploidization. (From Wolfe and Shields 1997.)

Although polyploidization events are generally rare in animals, the ray-finned fish lineage appears to be exceptional. The discovery that the zebrafish has seven Hox clusters, compared with four in tetrapods, led to the hypothesis that it is an ancient polyploid (Amores et al. 1998). More in-depth analyses of Hox and other genes in medaka and pufferfish suggest that the polyploidization event took place prior to the radiation of the ray-finned fish lineage (Taylor et al. 2001a, 2003; Christoffels et al. 2004; Jaillon et al. 2004; Crow et al. 2006). Under this interpretation, the presence of seven Hox clusters in the zebrafish, rather than the predicted eight, implies the loss of one cluster following polyploidization. As in land plants, this basal polyploidization event was followed by numerous secondary genome duplications, including those at the bases of the lineages of salmonids (Johnson et al. 1987), suckers (Ferris and Whitt 1979), sturgeon (Ludwig et al. 2001), and carp (David et al. 2003).

Taking one step further back in the vertebrate phylogeny, the presence of four Hox clusters throughout the tetrapod lineage, as well as multiple copies of a number of other genes in tetrapods relative to invertebrates, suggests the occurrence of substantial gene duplication prior to the divergence of ray-finned fishes and tetrapods. Based on rather limited data and indirect inference, Ohno (1970) first launched the idea that two adjacent rounds of polyploidization preceded the emergence of the major lineages of jawed vertebrates. Subsequently christened the "2R" hypothesis (for "two rounds" of polyploidization), Ohno's conjecture implies that the ancestral jawed vertebrate was an octoploid. This idea has been extraordinarily difficult to test, largely because of the enormous time since the putative duplication events took place (450–550 MYA). Over such a long period, enough chromosomal rearrangements, gene removals, and secondary segmental duplications have taken place to all but obliterate most ancestral linkage groups, and the silent sites of protein-coding genes are so saturated with mutations that accurate

dating of divergence times between candidate ancestral duplicates is essentially impossible.

Hughes and Friedman (2003) have questioned the validity of the 2R hypothesis on several grounds. First, based on the assumption that one duplication yields two copies of each gene and a second then yields four, they suggest that the 2R hypothesis implies a 4:1 ratio of gene family sizes between vertebrates and invertebrates. In contrast, the peak ratio of human gene family sizes to those of *Drosophila* (or *Caenorhabditis*) is close to 1:1, although ratios as large as 4:1 are found in many cases. Even in a comparison of vertebrates with the invertebrate cephalochordate *Branchiostoma*, a ratio of 2:1 is more common than 3:1, which in turn is more common than 4:1 (Furlong and Holland 2002). One weakness of using such observations to reject the 2R hypothesis, however, is that the 4:1 ratio prediction ignores the ubiquitous turnover of duplicated genes. Even if the half-life of vertebrate duplicate genes were as long as 100 million years (see below), almost all duplicates arising from an event about 500 MYA would have been silenced, and many smaller-scale secondary duplications would also have occurred. Thus, ratios of tetrapod:invertebrate gene family sizes are not particularly informative with respect to the 2R hypothesis, nor do they provide grounds for ruling out alternative hypotheses, such as ongoing segmental duplication.

Second, Hughes and Friedman (2003), as well as Martin (1999), argue that the 2R hypothesis predicts an (AB)(CD) topology for the genealogical trees of four-member families of vertebrate genes, with the AB and CD pairs of genes representing paralogs from the most recent genome duplication event and the (AB) versus (CD) clades representing the first duplication event (Figure 8.7). Although well over half of the estimated gene genealogies deviate from the (AB)(CD) expectation, there are again substantial reasons to question the informativeness of such an analysis. First, a large fraction of phylogenies with deep and contiguous internal nodes can be expected to deviate from the predicted pattern due to errors in phylogenetic construction alone. Second, with secondary duplications, rearrangements, and gene silencings overlaid on the original polyploidization events, a wide array of tree topologies is expected (recall, for example, Figure 1.5). Finally, if the original duplication events resulted from autopolyploidy, deviations from the (AB)(CD) pattern can arise naturally from allelic sorting prior to conversion of the octoploid state to functional diploidy (Furlong and Holland 2002). Thus, the failure to consistently recover an (AB)(CD) topology for vertebrate genes with four family members has little power for testing the 2R hypothesis.

Despite these uncertainties about the validity of the 2R hypothesis, there is little question that considerable gene duplication of some sort occurred deep in the chordate lineage. For example, after factoring out the large number of fairly recent duplications in the human genome, Panopoulou et al. (2003) dated a peak of ancient vertebrate duplicate genes at about 500 MYA. Similarly, by performing phylogenetic analyses on putative paralogous link-

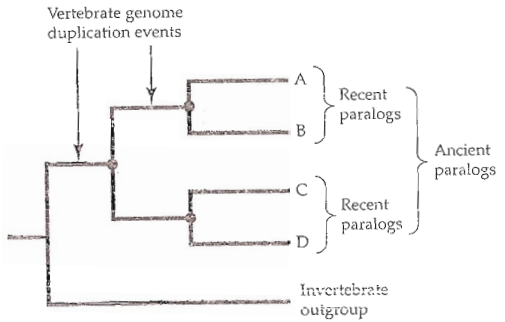

Figure 8.7 Idealized tree of relatedness of the members of a vertebrate gene family under the 2R hypothesis. The gray area denotes a set of genes within a single vertebrate genome resulting from two prior genome duplication events. A single copy of the gene (black dot) exists in the ancestral vertebrate as well as in the invertebrate outgroup. In the first round of polyploidization, this gene is duplicated to form the paralogous copies denoted by the green and blue dots. A second round of polyploidization then gives rise to four copies (A and B in the green lineage, and C and D in the blue lineage). As discussed in the text, this (AB)(CD) expectation requires that both copies survive the initial polyploidization event, and it also ignores the numerous ways in which secondary duplications and gene losses can lead to the same tree structure regardless of whether the 2R hypothesis is correct.

age groups using *Drosophila* and *Caenorhabditis* as outgroups, McLysaght et al. (2002) found a peak in the age distribution of duplicate genes prior to the emergence of vertebrates but subsequent to the protostome–deuterostome split. Additional analyses suggest that this peak postdates the split between cephalochordates and vertebrates (Figure 8.5) but predates the origin of jawed vertebrates (Abi-Rached et al. 2002; Escriva et al. 2002; Furlong and Holland 2002; Gu et al. 2002a; Robinson-Rechavi et al. 2004). Such timing is consistent with Ohno's conjecture that a massive phase of gene duplication provided the genetic substrate deployed in the subsequent emergence of the unique morphological innovations of jawed vertebrates (see Figure 8.5). Whether such extensive duplication activity involved one, two, or even three polyploidization events or was simply a period of substantial segmental duplication remains uncertain, but the most thorough study to date, a joint analysis of gene locations and family-member phylogenies, supports the 2R hypothesis (Dehal and Boore 2005). The unresolved matter of whether amplification of gene number had a causal role in phenotypic diversification is another issue entirely.

Finally, it is notable that genes arising via polyploidization generally exhibit enhanced longevity relative to those arising by segmental duplications. For example, based on the results cited above for yeast, the half-life

of a duplicate gene arising by a localized duplication event is about 17×10^6 generations, whereas that for a duplicate arising via the ancient polyploidization event is about 27×10^6 years (and presumably many more generations) (Table 8.2). Similarly, vertebrate genes arising by segmental duplication have half-lives of about 2×10^6 years, whereas those arising by polyploidization in ray-finned fishes have average half-lives of about 45×10^6 years. Finally, the half-lives of genes arising by polyploidization in plants appear to be approximately tenfold greater than those arising by small-scale events.

One possible explanation for the greater longevity of duplicates arising via polyploidization involves dosage requirements. When a complete genome is duplicated, each gene's expression is expected to remain in the same stoichiometric relationship with all of its interacting partners, a balance that may be favored by selection. In contrast, when a single gene is duplicated, it is immediately out of balance with its partners, potentially leading to functional difficulties. This hypothesis is consistent with two observations in yeast: first, that duplicate genes whose products participate in protein complexes are relatively underrepresented among duplications with segmental origins; and second, that members of interacting pairs of genes tend to be co-duplicated (Papp et al. 2003a). The enhanced longevity of duplicate genes derived from polyploidization may also be facilitated by the complete conservation of surrounding regulatory sites and the ancestral pattern of gene expression at the time of duplication.

Table 8.2 Proportions of genes surviving a polyploidization event

PHYLOGENETIC GROUP	FRACTION SURVIVING	DATE OF EVENT (MYA)	HALF-LIFE (MY)	REFERENCE
Animals				
Catastomids	0.50	50	50	Ferris and Whitt 1979
Cyprinus carpio (carp)	0.60	12	16	David et al. 2003
Loaches	0.25	28	14	Ferris and Whitt 1977
Salmonids	0.50	100	100	Allendorf et al. 1975
Xenopus laevis (frog)	0.77	30	80	Hughes and Hughes 1993
Land plants				
Arabidopsis thaliana	0.33	50	31	Ermolaeva et al. 2003
Oryza sativa (rice)	0.21	70	31	Paterson et al. 2004
Zea mays	0.72	11	23	Gaut and Doebley 1997
Yeast				
Saccharomyces cerevisiae	0.08	100	27	Wolfe and Shields 1997

Mechanisms for the Preservation of Duplicate Genes

To be successful in the long term, a duplicate gene must first drift toward fixation, and then, once it has risen to a high frequency, the selective forces for its maintenance must be sufficiently large to prevent its subsequent loss by degenerative mutation. The preceding results indicate that the vast majority of gene duplicates arising by segmental duplication experience an early exit from the population, most probably failing to ever reach fixation. However, a minority of duplicates can be retained for long periods. The precise mechanisms by which duplicate genes are preserved, three of which are discussed below, have a fundamental bearing on genome evolution. For example, the reciprocal preservation of both members of a pair of duplicates leads to an expansion in genome size, while the preservation of a new unlinked duplicate combined with the loss of the ancestral copy has no effect on gene number, but does induce an alteration of the genetic map.

Neofunctionalization

One of the more notable mechanisms for the preservation of a pair of gene duplicates is the process of neofunctionalization, whereby one copy acquires a beneficial mutation that results in a new function. Models of neofunctionalization via gene duplication generally assume that new beneficial functions are acquired at the expense of essential ancestral functions, the unspoken reasoning being that selectively advantageous mutations with no negative pleiotropic effects on wild-type fitness should have had no barriers to fixation prior to duplication. Under this reasoning, the temporary phase of redundancy provided by gene duplication is thought to release one copy from prior selective constraints, thereby enabling it to take on a previously forbidden adaptive feature (e.g., Ohno 1970). Although it is frequently assumed that the newly arisen copy will be the recipient of a new function, natural selection makes no decision as to which copy to tinker with. Instead, the early trajectories of the members of a duplicate pair are defined largely by their allelic ancestry or by early chance mutational events that occur in one copy or the other. The simplest version of this model, which assumes the duplication to be initially fixed in the population, extends back at least to Haldane (1933) and was explored quantitatively by Walsh (1995). In most of the following discussion, however, we will consider the more realistic case in which a duplicate arises as a single copy in a single member of the population, as the early phase of establishment can be critical.

Although most theoretical treatments of neofunctionalization have focused on mutations arising after the duplication event, Spofford (1969) made the key observation that the process need not await the arrival of new mutations. Her reasoning was based on the simple fact that the spectrum of mutations arising after a duplication event must be the same as that of mutations arising prior to duplication. Under this view, prior to duplication, a

mutant allele endowed with a beneficial function at the expense of an essential ancestral feature may be maintained at low frequency by balancing selection, provided that heterozygotes have a selective advantage (s) relative to wild-type homozygotes. Such ancestral polymorphisms can facilitate the route to neofunctionalization in two ways (Figure 8.8). First, if the duplicate locus is founded by a neofunctionalized allele, its fixation can be promoted by positive selection while the original locus retains the ancestral function. Second, if the duplicate locus is founded by a "wild-type" allele, which drifts by chance to a high enough frequency, the selective regime at the ancestral locus will be altered from balancing selection to positive selection for the previously low-frequency neofunctional allele. In either case, the final outcome is functionally equivalent to the fixation of heterozygosity, with one locus becoming essentially monomorphic for the wild-type allele and the other for the neofunctionalized allele.

A striking example of this process involves the evolution of insecticide resistance in the mosquito *Culex pipiens* (Lenormand et al. 1998). The acetyl-

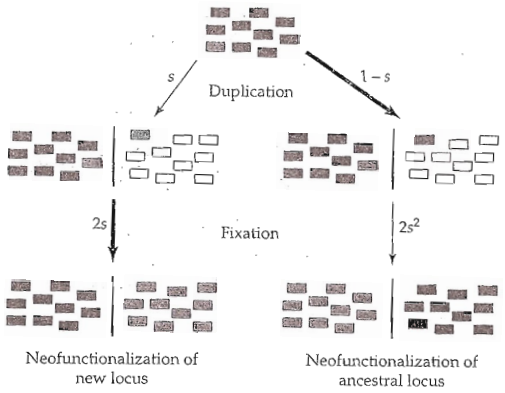

Figure 8.8 Neofunctionalization following duplication of an ancestral locus carrying an adaptive polymorphism. The individual boxes denote alleles in the gene pool. Prior to duplication, the green (neofunctional) allele is unable to go to fixation because, despite its advantage in heterozygotes, it is lethal in the homozygous state due to the absence of an essential wild-type function. The duplication event begins with either the wild-type (red) allele or the neofunctional (green) allele, with the relative frequencies of the starting states (denoted by arrows with different thicknesses and the algebraic expressions along the arrows) equal to the allele frequencies at the ancestral locus. Horizontal lines separate genes at the two loci; "absentee" alleles at the time of the duplication are denoted by hollow boxes. Although the path on the left is less likely at the duplication stage, once initiated, it is more likely to go to fixation because of the initial elevated advantage of the green allele at the new locus.

cholinesterase enzyme in this species normally plays an essential role in the central nervous system, but a mutant allele at the locus also confers resistance to organophosphate insecticides. Because the mutant allele reduces fitness in the absence of insecticides (Berticat et al. 2002), it is maintained at a very low frequency in insecticide-free environments. However, when mosquitoes were challenged with organophosphates, a linked combination of a wild-type and a resistant-type allele rose rapidly to a high frequency, presumably because the first member of the pair provided the essential ancestral gene functions while the second ameliorated the toxicity of the environment. The linked pair of duplicates may have been present in the absence of insecticides, but kept at a very low frequency by the negative effects of increased dosage. Similarly, trichromatic vision in some primates appears to have become established by the parallel fixation of preexisting opsin alleles following the duplication of an ancestral locus (Jacobs et al. 1996; Dulai et al. 1999).

The conditions necessary for the maintenance of neofunctional alleles at an ancestral locus by balancing selection have been worked out (Lynch et al. 2001). The key requirement is that the power of random genetic drift be sufficiently weak relative to the strength of balancing selection (approximately $1/N_g < s^2/8$, where N_g is the effective number of genes per locus—equivalent to the effective population size for a haploid and approximately twice that for an outcrossing diploid; see Chapter 4). Provided this condition is met, the neofunctional allele will be present at the ancestral locus with approximate expected frequency s (e.g., an allele that is lethal in the homozygous state but increases fitness by 1% in the heterozygous state would have an expected frequency of 1%, provided that $N_g > 80{,}000$). Thus, a mild heterozygote advantage combined with a moderately large effective population size is sufficient to poise a population for progression toward neofunctionalization following a duplication event, eliminating the waiting time for new mutations.

The probability of preservation of a pair of duplicate genes by neofunctionalization depends on several additional factors (Lynch et al. 2001; Walsh 2003). In large populations, the probability of neofunctionalization increases with s^2, provided the duplicate loci are unlinked. This scaling can be understood most easily by noting that the new locus will be founded by a wild-type allele with probability $(1 - s)$ and by a neofunctionalized allele with probability s (see Figure 8.8). In the former case, the founder allele will have a selective advantage defined by the frequency of neofunctional homozygotes at the ancestral locus (s^2), because "absentee" homozygotes at the new locus are lethal on this genetic background. In the latter case, the marginal selective advantage of the founder allele over the absentee allele is simply the selective advantage s. The probability of neofunctionalization by each of these paths is equal to the product of the probability of the starting condition and the fixation probability (which, in a large population, is approximately equal to twice the selective advantage; see Chapter 4). Assuming small s so that $(1 - s) \approx 1$, each path has a probability of approximately $2s^2$.

This overall neofunctionalization probability is actually an upper bound, as it assumes the duplicate loci to be unlinked. This point can most easily be understood by considering a pair of tandemly linked duplicates in which both members have the same initial allelic state. In the extreme case of complete linkage, a tandem pair of neofunctionalized alleles cannot proceed to fixation, as this would lead to the complete loss of the essential ancestral function.

In this latter case, as well as in the more general case in which $N_g < 8/s^2$ and the neofunctional allele cannot be maintained at the ancestral locus by selection, permanent neofunctionalization by gene duplication requires a starting point at which both loci harbor only wild-type alleles. Such a condition imposes a particularly precarious phase of initial establishment, making it unlikely that the probability of neofunctionalization will exceed $1/(2N)$, the initial frequency of the active allele at the new locus, unless there is an intrinsic selective premium on the duplication itself (e.g., a dosage advantage). If the duplication is initially neutral, it will drift to fixation with probability $1/(2N)$, and its subsequent fate will be determined by the relative rates of fixation of inactivating versus preservational mutations. Given an assumed initial state of neutrality, the rate of inactivation is simply equal to the degenerative mutation rate, whereas the rate of preservation by neofunctionalizing mutations is equal to the rate of origin of such mutations times the probability of fixation (necessarily less than 1). Thus, because we generally expect most mutations to be deleterious, the probability of neofunctionalization in the absence of neofunctional alleles in the base population is likely to be just a small fraction of $1/(2N)$ (Walsh 1995).

On the other hand, even in the absence of preexisting neofunctional alleles, if there is an immediate selective advantage for a duplication, as would be the case if a larger amount of the gene product were beneficial (Zhang 2003; Francino 2005; Kondrashov and Kondrashov 2006), then the probability of the first (preservational) step toward neofunctionalization can be greater than $1/(2N)$. As noted in Chapter 4, positive selection will be effective only if the intrinsic advantage of a duplicate (s_d) is sufficiently greater than the power of random genetic drift (i.e., $2N_g s_d > 1$). Moreover, after a duplicate has been preserved by this process, neofunctionalization will follow only if the origin of a new function at one locus provides a benefit (s) that significantly exceeds the initial advantage associated with gene dosage [i.e., $2N_g(s - s_d) > 1$]. All of these observations reinforce the general principle that the neofunctionalization of a duplicate gene is a large-population phenomenon.

Finally, under all of the theory discussed above, neofunctionalization evolves through the progressive modification of an active allele. Ohno (1970) suggested an alternative scenario by which a silenced gene duplicate (i.e., a transient pseudogene) might provide the substrate for the origin of a novel function (see also Marshall et al. 1994). By accumulating a series of molecular changes in a neutral fashion, an inactivated locus might eventually yield a beneficial product that would be impossible to acquire via natural selection alone. Imagine, for example, a silent allele at a duplicated locus separated from a beneficial state by intermediate mutational steps associated with low fitness (if expressed). If, after the fortuitous acquisition of mutations with jointly beneficial effects, such an allele were to be reactivated somehow, it would be strongly favored by natural selection, possibly displacing the ancestral locus (in the absence of negative pleiotropy) or coexisting permanently with it (in the presence of a trade-off). In principle, this mechanism for vaulting an adaptive valley might be facilitated by gene conversion between the silenced and expressed loci (Hansen et al. 2000). To date, however, there is no evidence that these sorts of Lazarus effects play an important role in evolution.

The masking effect of duplicate genes

Because all loci harbor suboptimal alleles due to the recurrent introduction of deleterious mutations, it is sometimes thought that duplicate genes have an intrinsic selective advantage associated with their ability to mask the effects of deleterious mutations at the ancestral locus. But is the magnitude of such an indirect advantage great enough to promote the permanent preservation of duplicate genes? Fisher (1935) realized that even in an effectively infinite population (in which the efficiency of selection is maximized), two genes with identical functions will not be mutually maintained by this process unless their mutation rates to defective alleles are identical. If this is not the case, the gene with the higher mutation rate will eventually be silenced by the differential accumulation of genetic load. In principle, permanent retention of duplicate genes might be achieved by a delicate balance between differential selective advantages and differential mutation rates of a pair of duplicate loci (e.g., with one gene operating more efficiently and the other having a lower mutation rate to null alleles) (Nowak et al. 1997), but the conditions required under this model are unrealistically stringent.

In finite populations, not even identical mutation rates are sufficient for the permanent retention of duplicate genes via the masking effect (Clark 1994; Lynch et al. 2001; O'Hely 2006). Consider, for example, a segregating recessive lethal allele, which in a large population has an equilibrium homozygote frequency at a single-copy locus equal to the null mutation rate, μ_c (Crow and Kimura 1970). Under the masking model, μ_c is equal to the initial selective advantage of an otherwise redundant functional duplicate at a new locus, as the duplicate modifies fitness only in individuals that have no functional gene at the ancestral locus. However, because μ_c is also the rate of silencing of genes at the duplicate locus, these two factors cancel exactly, rendering the duplicate effectively neutral and vulnerable to eventual loss by random genetic drift. If null alleles have fitness effects in heterozygotes, a duplicate gene can potentially mask the effects of suboptimal genotypes in a larger fraction of the population, but even in this case, the net advantage of the masking effect is on the order of the mutation rate.

A deleterious allele causing a reduction of fitness in heterozygotes of hs has an expected frequency $q = \mu_c/(hs)$ under selection–mutation balance, so with a fraction $2(1-q)q \approx 2q$ of the population heterozygous at the original locus, the selective advantage of a newly arisen duplicate is just $2q \cdot hs = 2\mu_c$ (Otto and Yong 2002; Proulx and Phillips 2005). After subtracting out the rate of loss of the duplicate gene by mutation, and recalling the definition of effective neutrality (see Chapter 4), we see that the condition for the maintenance of a duplicate gene via the masking effect is $2N_g\mu_c > 1$. In other words, the effective gene number per locus must exceed $1/(2\mu_c)$.

Although the preceding arguments raise significant questions about the sufficiency of the power of the masking effect to maintain duplicate genes, they do not rule out the possibility that duplicates maintained by other processes still have a buffering potential. However, although genetic redundancy almost certainly masks some deleterious mutational effects, quantitative information on the matter is scant. In the yeast *S. cerevisiae* and the nematode *C. elegans*, the fitness consequences of knockouts of single members of pairs of duplicate genes are smaller than those of knockouts of single-copy genes, and the effects of knockouts increase with the magnitude of sequence divergence between paralogs (Gu et al. 2003; Conant and Wagner 2004). Taken at face value, such observations suggest a diminishing incidence of overlapping functions as a duplicate pair ages. However, studies of extant duplicates suffer from ascertainment bias. In yeast, for example, those genes that have the smallest knockout effects when present in a single copy are the most likely to exist as duplicates (He and Zhang 2005a; Prachumwat and Li 2006), so the small effects of deletions of single members of such pairs are not entirely due to a buffering effect, but at least in part a simple consequence of their nonessentiality.

Finally, it should be noted that the masking effects described above are concerned with compensation for mutationally silenced alleles, whereas duplicate genes might also provide a buffer against transient cellular mishaps causing localized absence of gene expression by normally active genes (Tautz 1992; Nowak et al. 1997). The potential sources of such developmental errors include somatic mutations and errors in transcription and translation. Following the logic outlined above, for this masking mechanism to maintain a duplicate gene by natural selection, the fitness consequences of developmental errors at a locus would have to exceed the rate of origin of null mutations at the duplicate locus, and this net difference, in turn, would have to exceed the power of random genetic drift.

In summary, despite their seductive nature, the various masking models for the preservation of duplicate genes require rather special sets of mutational conditions and very large effective population sizes to enable the very weak selective advantages of redundancy to come to prominence. Perhaps the most serious challenge to the idea that masking plays a prominent role in duplicate-gene retention is the general paucity of duplicate genes in prokaryotes despite their haploid nature and exceptionally high N_g.

Subfunctionalization

Given the difficulties with the various masking hypotheses, neofunctionalization has often been assumed to be the *only* mechanism that can permanently preserve duplicate genes. Under this view, however, the vast majority of new gene duplicates are expected to be lost within a relatively short time because of the rarity of neofunctionalizing relative to nonfunctionalizing mutations. In the absence of positive selection, a fraction $[1 - 1/(2N)]$ of newly arisen gene duplicates will be lost by random genetic drift in an average of just $2\ln(N_g)$ generations (Kimura and Ohta 1969), a flash on the evolutionary time scale. Moreover, the small remaining fraction, $1/(2N)$, that manages to drift to fixation is also expected to fall victim to silencing mutations relatively quickly. If $N_g\mu_c << 1$, the average time to silencing of a fixed duplicate is approximately equal to the mean waiting time for the appearance of a null mutation at one of the loci, $1/(2\mu_c)$ generations. On the other hand, when $N_g\mu_c >> 1$, null mutations are common, and the time to silencing depends largely on the time required for one of them to drift to fixation at one of the loci, about $2N_g$ generations (see Chapter 4), with the slight initial masking advantage of the new allele (described above) prolonging its survival up to $5N_g$ generations (Watterson 1983; Lynch and Force 2000b).

Although these predictions of a relatively rapid demise of the vast majority of duplicate genes are in rough accord with the evolutionary demographic analyses reported above for segmental duplications, they are inconsistent with the large levels of duplicate-gene retention in ancient polyploids (see Table 8.2). Moreover, the lack of evidence that polyploid lineages have evolved unusually large numbers of new gene functions suggests the involvement of preservational mechanisms other than neofunctionalization. As noted above, positive selection associated with dosage requirements may play a role in polyploid lineages, but a more general mechanism for duplicate-gene preservation, relevant even to segmental duplications, derives from a broader view of gene structure than assumed under the classic model.

Many eukaryotic genes, particularly those in multicellular species, have complex, modular regulatory regions, alternative splicing mechanisms, and/or functional domain structures. Such genes are naturally endowed with independently mutable subfunctions, in the sense that mutations that cause the loss of function in one particular tissue or developmental period do not necessarily affect other tissue- or timing-specific aspects of their expression. The widespread existence of such structural complexities leads to the prediction that duplicate-gene preservation will sometimes result from the partitioning of ancestral gene functions through complementary loss-of-function mutations in paralogous copies (Figure 8.9). Under the DDC (Duplication–Degeneration–Complementation) model of Force et al. (1999), subfunctionalization is driven entirely by degenerative mutations, which we know to be much more common than beneficial mutations (Lynch et al. 1999). Special cases of this model have also been discussed by Hughes (1994) and Stoltzfus (1999), and Taylor and Raes (2004) cite earlier relevant references.

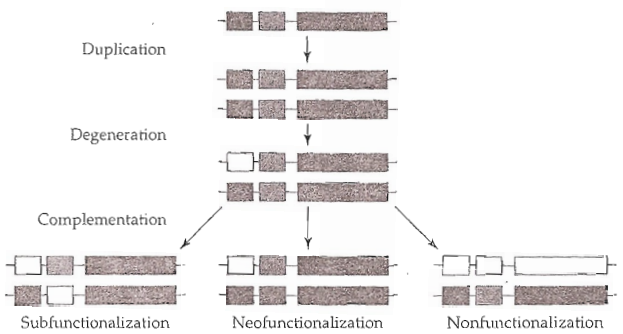

Figure 8.9 The DDC model for the alternative fates of duplicate genes. The ancestral gene is depicted as having two independently mutable regulatory regions (one blue and one green), each driving expression in a particular tissue or developmental period. Solid boxes denote fully functional regulatory and coding regions, whereas open boxes denote loss of function, and the red box denotes the gain of a new beneficial function. Each pair of genes reflects the fixed haploid state of the population. Following the duplication event, the first degenerative mutation eliminates a subfunction of one of the copies. The second mutational event then dictates the final fate of the pair—subfunctionalization, with the second copy acquiring a complementary loss-of-subfunction mutation; neofunctionalization, with the second copy acquiring a novel, beneficial expression pattern at the expense of an ancestral subfunction; or nonfunctionalization, with the first copy losing all functional ability. (From Force et al. 1999.)

Duplicate-gene preservation by subfunctionalization is a two-step process. First, one of the genes becomes fixed for a mutation that eliminates an essential subfunction, permanently preserving the second copy. Loss of an alternative subfunction by the second copy then reciprocally preserves the first copy. If the effective population size is sufficiently small that segregating null mutations are typically rare ($N_g\mu_c << 1$), the fate of a newly arisen duplicate gene under the DDC model depends almost entirely on the relative rates of origin of mutations that abrogate single subfunctions (μ_r) versus those that cause complete nonfunctionalization (μ_c), and the probability of subfunctionalization can be approximated with combinatorial logic. Consider, for example, a gene with two independently mutable subfunctions, with the rate of mutation to new beneficial functions being of negligible importance. The probability that a newborn duplicate will drift to fixation is its initial frequency, $1/(2N)$, and having arrived at that point, the probability that the first fixed mutation will eliminate a subfunction from one of the genes is simply the fraction of mutations that are of the subfunctionalizing type, $2\mu_r/(2\mu_r + \mu_c)$. After fixation of a subfunctionalized

allele at one locus, the intact locus is no longer free to lose its now unique subfunction, so it accepts mutations at rate μ_r, whereas the partially debilitated locus is free to become completely silenced (at rate $\mu_r + \mu_c$). Thus, the fraction of permissible mutations that eliminate the complementary subfunction at the intact locus in the second step is $\mu_r/(2\mu_r + \mu_c)$. The probability of subfunctionalization (P_{sub}) is equal to the product of these three probabilities, $\alpha^2/(4N)$, where $\alpha = 2\mu_r/(2\mu_r + \mu_c)$ is the fraction of degenerative mutations that eliminate single subfunctions. Under this two-subfunction model, P_{sub} approaches a maximum of $1/(4N)$ as α approaches 1 (all mutations are of the subfunctionalizing type) because there is a 50% probability that the first two mutations will be incurred by the same copy (leading to nonfunctionalization) and a 50% probability that the two copies will incur complementary mutations.

A number of factors can increase the probability of subfunctionalization above $1/(4N)$, but none of them is likely to move the upper bound beyond $1/(2N)$, the probability of initial fixation of an entirely neutral duplication. For example, increasing the number of independently mutable subfunctions increases P_{sub} by increasing the number of paths by which complementary loss-of-function mutations can be acquired by the two copies (Lynch and Force 2000b). Mutations with partial effects on gene expression will also increase the probability of duplicate-gene preservation, even providing a retention mechanism for duplicates whose expression patterns cannot be subdivided spatially or temporally in development (Lynch and Force 2000b; Duarte et al. 2006). Such preservation, known as quantitative subfunctionalization, occurs whenever the total capacity of both loci is degraded to the extent that their joint presence is needed to fulfill the requirements of the single-copy ancestral gene (as in the case of duplicate enzyme genes that acquire activity-reducing mutations). Consider, for example, a gene with a single function, and let n be the number of mutations with partial effects necessary to completely eliminate its function. If μ_p is the total rate of origin of such mutations and $\rho = \mu_p/(\mu_p + \mu_c)$ is the fraction of the total pool of mutations with partial effects, then the upper limit to P_{sub} under this model, $\rho^2/(2N)$, is approached as the average effects of partially debilitating mutations decline to zero (i.e., as n approaches infinity). This limit approaches a maximum value of $1/(2N)$ as mutations with partial effects become more predominant (i.e., as ρ approaches 1). Finally, as noted above, newly arisen duplicates need not always be complete, and if critical regulatory elements are missing from the flanking regions at the time of the duplication event, the first step toward subfunctionalization may be fulfilled at the outset (Averof et al. 1996). For a gene with two subfunctions, the loss of one subfunction at the time of duplication would increase the probability of subfunctionalization from $\alpha^2/(4N)$ to $\alpha/(4N)$.

The subfunctionalization model makes unique predictions about the scaling of the probability of duplicate-gene preservation with population size (Box 8.1). Assuming that there is no initial selective consequence of subfunctionalization, duplicate genes should be preserved more often by this

mechanism in small populations and somewhat more commonly when they are tandemly linked. These predictions contrast with those of the neofunctionalization model, which predicts that duplicate genes should be preserved more often in large populations and more commonly when they are unlinked. These divergent expectations have a potentially important bearing on our understanding of the mechanisms leading to genome expansion. If neofunctionalization is the predominant mechanism of duplicate-gene preservation, then larger populations, which harbor more targets for rare beneficial mutations, should experience a greater expansion of gene number (provided the duplication rate itself is comparable, and factoring out quantum changes due to polyploidization). Noting that the mouse lineage has

Box 8.1 Effective Population Size, Linkage, and the Probability of Subfunctionalization

The expressions derived in the text for the probability of subfunctionalization assume a sufficiently small effective population size that (1) each step in the process proceeds to completion before the next key mutational event occurs and (2) differences in the mutational vulnerabilities of alternative alleles have negligible selective consequences. These conditions, which are generally met if the effective population size is small enough that $N_g\mu_c << 1$ (where μ_c is the rate of nonfunctionalizing mutation per generation), result in a situation in which the rate of subfunctionalization at a locus is essentially independent of both the population size and the degree of linkage between duplicates. Assuming diploidy, the population-level rate of duplication is $2NB$, where N is the population size and B is the physical rate of gene duplication and, conditional on duplication, the probability of subfunctionalization is $\alpha^2/(4N)$, where α is the fraction of degenerative mutations that eliminate single subfunctions. The rate of subfunctionalization is the product of these two components, $B\alpha^2/2$ (Lynch and Force 2000b).

In larger populations, both population size and degree of linkage play important roles in determining the probability of subfunctionalization (Lynch et al. 2001). Specifically, for $N_g\mu_c > 1$ and unlinked duplicates, P_{sub} approaches zero at large N_g because during the first phase of the process (the $\sim 2N_g$ generations required for a newly arisen duplicate to drift to initial fixation), essentially all descendants of the initial duplicate will acquire silencing mutations.

The situation is slightly more complicated for linked duplicates. The probability of subfunctionalization again declines to zero at large N_g, but at a somewhat higher threshold value of N_g than for unlinked duplicates. A linked pair of duplicates initially has a weak selective advantage (approximately equal to μ_c) over a single-copy gene, resulting from the fact that complete inactivation of a "two-copy" allele requires the silencing of both members of the pair. This advantage gives a linked pair of duplicates a slight boost in the initial fixation process relative to the neutral expectation of $1/(2N)$. However, once such a pair becomes subfunctionalized, the tables are turned: because such a linked pair requires two coding regions to carry out the ancestral subfunctions, it is a larger mutational target than a fully functional single-copy allele, and therefore has a permanent selective disadvantage (again equal to μ_c). These types of insights would not have been possible without a formal population genetic analysis.

experienced roughly 60% more gains and 45% fewer losses of duplicate genes established prior to the rodent–primate divergence than has the great ape lineage (assumed to have a smaller long-term effective population size), Shiu et al. (2006) have argued that positive selection plays a more important role in duplicate-gene preservation than does degenerative mutation. However, because the generation length of rodents is substantially shorter than that of primates, on a per-generation basis (the relevant scale for genomic evolution), there are actually substantially more gains and losses of duplicates in the human lineage (approximately 6 and 19 times more, respectively, if the average generation time leading to humans is assumed to be ten times that for rodents). Thus, these comparative analyses of mammalian duplicates actually support a central role for subfunctionalizing and nonfunctionalizing mutations, as do numerous other observations to be discussed below.

By postulating a preservational process driven entirely by degenerative mutations, the subfunctionalization model provides a null hypothesis for the interpretation of patterns of survival of duplicate genes. Subfunctionalization, however, may be the beginning, not the end, of new evolutionary pathways. Consider, for example, a single-copy locus that is a victim of a "jack of all trades, master of none" syndrome, such that an adaptive conflict exists between its subfunctions. If such a gene is duplicated, complementary loss-of-subfunction mutations are expected to alter the selective landscape experienced by the two members of the duplicate pair, enabling each copy to become more refined to its specific subset of tasks (Piatigorsky and Wistow 1991; Hughes 1994) and perhaps opening up previously unavailable pathways to neofunctionalization. By this means, two of the most common forms of genomic upheaval, gene duplication and degenerative mutation, may provide a unique mechanism for the creation of novel evolutionary opportunities through the elimination of pleiotropic constraints.

The Fates of Duplicated Protein Sequences

The alternative models for the maintenance of duplicate genes motivate several questions. First, do duplicate genes experience an early phase of relaxed purifying selection? Second, do functional novelties in duplicate genes arise out of preexisting polymorphisms (the balancing-selection model), as refinements of preexisting subfunctions in multifunctional single-copy genes (the adaptive-conflict model), or as de novo modifications following duplication? Third, are the mutational events that drive the critical phase of early preservation typically beneficial (neofunctionalization model) or degenerative (subfunctionalization model), and how frequently do such mutations influence regulatory versus coding sequences? Fourth, are the rates of molecular evolution in the two copies equal or asymmetrical (the latter being the expectation if just one copy embarks on an exploratory evolutionary pathway)? Answering these questions has been a major challenge.

The presentation in the previous section indicates that the relative incidence of nonfunctionalizing, subfunctionalizing, and neofunctionalizing

mutations should be a fundamental determinant of the fates of duplicate genes, but direct information on this matter is lacking. Thus, most attempts to understand the evolution of duplicate genes have resorted to a more indirect approach: comparative analysis of paralogous sequences in extant species. The cumulative nature of gene evolution substantially restricts the information that can be gleaned from such studies. In order to distinguish between the mutational events that lead to duplicate-gene establishment versus subsequent divergence, one would like to follow the historical record of substitutional changes incurred by both members of the pair from the time of origin to the time of preservation or elimination, but comparative analysis reveals only the end products of evolution. In addition, due to our poor understanding of the functional significance of noncoding regions of genes, almost all research in this area has been restricted to coding region evolution.

As noted in previous chapters, contrasts in the number of amino acid replacement substitutions per replacement site (R) and the number of silent substitutions per silent site (S) can provide crude information on the average form of selection operating on a pair of genes: an R/S ratio smaller than one implies that selection on replacement sites is predominantly of a purifying form, while a ratio greater than one implies directional selection for a change in function in one or both sequences. This criterion for distinguishing between directional and purifying selection is actually somewhat ambiguous in that an R/S ratio smaller than one averaged over the full length of a coding sequence does not preclude the possibility of strong adaptive divergence in one region of the protein on a more general background of purifying selection. In addition, an R/S close to one, which is consistent with neutrality, could also arise if one member of a pair were under purifying selection and the other under positive selection. On the other hand, a gene-wide estimate of R/S greater than one provides essentially unambiguous evidence of directional selection in one or both pair members. A deeper problem is that the estimates of R and S for any pair of extant duplicates are cumulative outcomes of the joint evolutionary pressures operating on both loci since the initial duplication event. In a simple two-gene analysis, a brief early phase of relaxed selection ($R/S \approx 1.0$) could easily be obscured by a subsequent prolonged phase of purifying or positive selection.

Patterns of molecular evolution

Some information on the temporal dynamics of selection on gene duplicates can be gleaned from observations on the joint distribution of R and S for the entire assemblage of gene pairs within a species, from the youngest newborn cohort to the less abundant ancient pairs. Although such an approach is unable to reveal the historical order of the mutational events associated with any particular gene pair, it may help reveal the average temporal pattern of selection experienced by all cohorts of duplicates. If, for example, the intensity of purifying selection operating on duplicate genes typically

increases with the age of a pair, this pattern should be reflected in a reduction in R/S in cohorts of increasing age, whereas the opposite is expected if selection is progressively relaxed.

One way to achieve insight into this matter is to assume a particular pattern in the way in which the instantaneous ratio of replacement to silent mutation rates (dR/dS) changes through time to yield a cumulative relationship between observed R values in gene pairs with different values of S (Box 8.2; Figure 8.10). From the observed behavior of cumulative R and S over the entire set of duplicates within a particular genome, it is then possible to indirectly infer the historical changes in average cohort-specific values of dR/dS that best explain the cumulative data. Fits of such a model to observations on complete genomic sets of gene duplicates from eukaryotes reveal fairly consistent patterns (Lynch and Conery 2000, 2003a), most notably a clear tendency for dR/dS to decrease with increasing S (Figure 8.11). The asymptotic values of dR/dS at low S for animals, yeasts, and land plants range from 0.36 to 1.00, with an overall average of 0.77 (implying that only about 23% of replacement substitutions are removed from relatively young duplicates by selection). In contrast, estimates of dR/dS at high S fall in the narrow range of 0.02–0.09, with an overall average of 0.05. Thus, whereas a substantial fraction of young duplicate pairs experience a phase of highly relaxed selection, by the time a duplicate pair has diverged about 10% at silent sites, reestablishment of a purifying selection regime has generally increased the stringency of selection against amino acid changes more than tenfold, to the point at which only about 5% of replacement mutations are able to proceed to fixation. Only a small fraction of intermediate-aged duplicates exhibit rates of replacement-site substitution in excess of the neutral expectation, no more than expected by chance in a large sample of multiple comparisons.

The conclusion that newly arisen gene duplicates generally experience an initial phase of relaxed selection is supported by a number of other studies. For example, the average cumulative R/S values for paralogs within a variety of prokaryotic and eukaryotic species are higher than those for single-copy genes separated by speciation events, but still virtually always less than 1.0 (Li 1985; Van de Peer et al. 2001; Kondrashov et al. 2002; Nembaware et al. 2002; Seoighe et al. 2003). Despite the consistency of these observations, numerous issues remain unresolved. In particular, elevated levels of dR/dS in young duplicates could be a consequence of relaxed selection against degenerative mutations at some sites, positive directional selection at others, or both. In addition, there may be an intrinsic bias to whole-genome analyses that lump all functional categories of genes together if there is significant variation in the duplicability of different gene categories. Heterogeneous patterns of gene duplicability do appear to be common. For example, protein-coding genes that evolve slowly and are expressed broadly when present in a single copy are preserved as duplicates more frequently than rapidly evolving or narrowly expressed protein-coding genes (Yang et al. 2003; Davis and Petrov 2004; Jordan et al. 2004; Brunet et al. 2006; Chap-

Box 8.2 Indirect Inference of Historical Patterns of Molecular Evolution in Duplicate Protein-Coding Genes

To account for the possibility that the ratio of instantaneous rates of replacement and silent substitutions (dR/dS) changes through time, a mathematical function such as that outlined in Figure 8.10A can be employed. This particular function allows for two different phases of divergence as well as a gradual transition between them. Assuming positive m, dR/dS starts with an expected value of $1/(a - b)$ at $S = 0$ (newly arisen duplicates) and declines to $1/a$ as S approaches infinity (ancient duplicates). (It is assumed here that both a and b are positive; a negative value of b would cause dR/dS to increase with S, but such behavior has not been seen with empirical data.) Other mathematical functions with more complex behavior can be constructed, but the specific model in Figure 8.10A is especially useful because it can be integrated to yield a simple algebraic expression for the historical development of the cumulative ratio of substitutions per replacement site and per silent site, as shown in Figure 8.10B. It is this cumulative R/S ratio that is observed in empirical studies, as opposed to the instantaneous ratio, which can vary from cohort to cohort.

On a log–log plot of R versus S, points with equal R/S ratios fall on a diagonal line, with the height of the line being defined by the magnitude of R/S. Thus, with the preceding model, a linear relationship appears between log S and log R when S is sufficiently small that $dR/dS \approx 1/(a - b)$. The response of log R to log S (the solid curve) then becomes shallower as a transition is made to a constant and lower ratio ($dR/dS \approx 1/a$). During a period in which genes are evolving in a neutral fashion, the response will converge with the diagonal describing $dR/dS = 1.0$ (the main diagonal in Figure 8.10B).

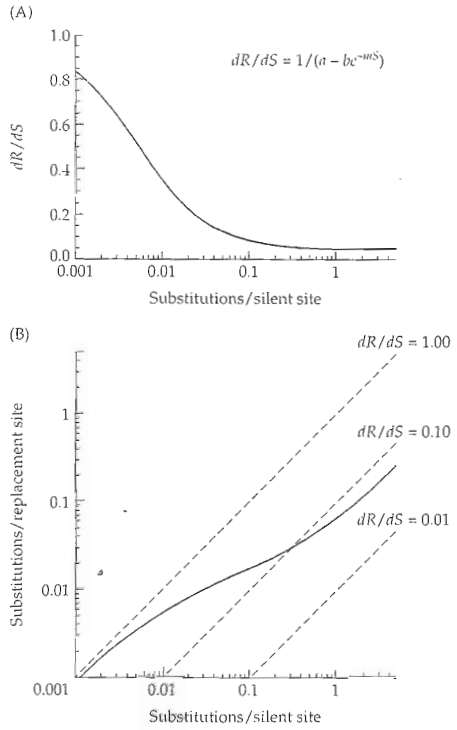

Figure 8.10 The change in the R/S ratio with increasing evolutionary time (measured in units of S). (A) The instantaneous ratio of replacement to silent substitutions (i.e., the ratio for mutations arising within the current cohort), defined by the inset equation with $a = 20$, $b = 19$, and $m = 10$. In this particular example, the ratio of replacement to silent substitutions initiates at 1.0 for newly arisen duplicates ($S = 0.0$) and gradually declines to a stable ratio of 0.05 as $S \rightarrow \infty$ (B) The cumulative behavior of R vs. S, obtained by integrating the equation in the top panel (see Equation 8 in Lynch and Conery 2003a). Here, the dashed lines represent points of equal R/S.

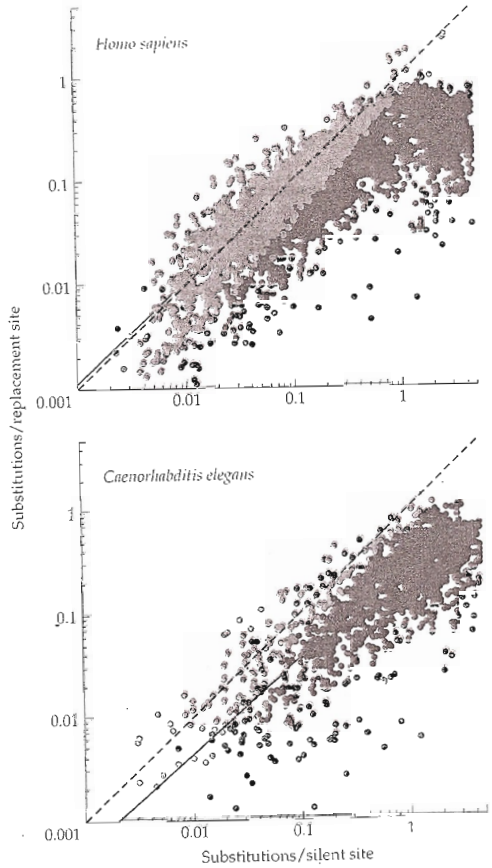

Figure 8.11 R versus S plots for duplicate gene pairs in two animal species. Red points denote pairs for which R is not significantly different from S (i.e., for which neutral evolution cannot be rejected). The diagonal dashed line denotes the neutral expectation, $R = S$, whereas the solid curve (mostly hidden behind the data points) is the fitted function described in the text. (From Lynch and Conery 2003a.)

man et al. 2006). The enhanced frequency of successful duplications of particular gene classes might be a consequence of their elevated susceptibility to subfunctionalization, a pattern that is consistent with large multidomain proteins having higher probabilities of preservation following polyploidization (Chapman et al. 2006). The elevated preservation of short genes with strong tissue-specific expression patterns following segmental duplication events (Urrutia and Hurst 2003) may be a simple consequence of an enhanced likelihood of being fully contained within a duplication span. Regardless of the underlying mechanism, the innate tendency of genes that are more prone to successful duplication to also have lower amino acid substitution rates implies that the decline in R/S in older duplicates seen in Figure 8.11 may be not only a consequence of an increase in the stringency of purifying selection on individual gene pairs over time, but also a reflection of the elevated extinction rate of duplicate pairs with high R.

Although the high values of R/S frequently observed in young cohorts of duplicate genes imply that mutations that never become fixed in single-copy genes are often able to accumulate in a nearly neutral fashion early after gene duplication, such observations do not address the common notion that just one member of a duplicate pair experiences an altered evolutionary trajectory after duplication. Resolution of this issue requires that both members of a paralogous pair be compared with an appropriate single-copy gene within an outgroup species; if both members of the pair are evolving under the same levels of constraint, they should be equally divergent from the outgroup sequence. In the first study of this kind, using mammalian outgroup sequences, Hughes and Hughes (1993) found no evidence for unequal rates of duplicate-gene evolution in the tetraploid frog *Xenopus laevis*, and similar results have been obtained with other taxa (Kondrashov et al. 2002; L. Zhang et al. 2002). Most of these analyses focused on relatively old duplicates, however, raising the possibility that an early phase of asymmetry was obscured by subsequent substitutional changes. Indeed, analysis of relatively young ($S < 0.3$) duplicate pairs from the human genome reveals a very high incidence (~60%) of asymmetrically evolving pairs (P. Zhang et al. 2003), and results from several other animals and yeasts also point to significant asymmetry in sequence divergence (Van de Peer et al. 2001; Conant and Wagner 2003; Chain and Evans 2006). The tendency for the more rapidly evolving members of human gene pairs to accumulate changes evenly across the molecule while the more slowly evolving copies exhibit spatially uneven substitution patterns is consistent with a relaxation of selection on the more rapidly evolving pair members (P. Zhang et al. 2003), as is the observation that the more slowly evolving members of yeast duplicate pairs tend to be more embedded in protein interaction networks and more critical to fitness (Kim and Yi 2006).

The elevated level of dR/dS, as well as the rate asymmetry, in young duplicate genes may simply be a consequence of chance events setting one copy on a trajectory of $dR/dS = 1.0$ until it is completely nonfunctionalized. However, the eventual return of surviving pairs of duplicate genes to a trajectory of low dR/dS is also consistent with the hypothesis that the early phase of relaxed selection frequently reflects the initial preservation of duplicates by subfunctionalizing degenerative mutations, as is the observed narrowing of gene-specific expression domains following duplication (described below). Single-copy vertebrate genes with more restricted tissue-specific patterns of expression are known to evolve more rapidly at the amino acid level, putatively because of a reduction in selective constraints (Hastings 1996; Duret and Mouchiroud 2000), so it is plausible that the early phase of high dR/dS in duplicate genes is a reflection of a narrowing of paralog expression patterns. In a broader evaluation of the possibility of subfunctionalization at the level of protein structure, Dermitzakis and Clark (2001) introduced a computational test for spatial variation in substitution patterns in the coding regions of paralogs and found that about half of mouse and human duplicate genes exhibit significant spatial variation among paralogous copies, some of which appears to be associated with functional domains.

Although these patterns are suggestive, statistical considerations raise significant caveats with respect to interpretations derived from observations on asymmetrical paralog divergence. As noted above, the fates of most duplicate genes are likely to be determined by the first few mutations incurred by one or both pair members. However, there is essentially no statistical power to detect significant asymmetry among paralogs until the average member has incurred a few dozen mutations (Lynch and Katju 2004). Moreover, from the standpoint of the masking model of duplicate-gene preservation, the question of interest is whether the two members of a pair have evolved in a more symmetrical pattern than expected by chance, not whether the evolutionary pattern is unbalanced. Testing for symmetry is even more difficult than testing for asymmetry. Even with an average of 100 to 300 substitutions per gene, it is virtually impossible to detect an exceptionally symmetrical rate of evolution across a pair of duplicates: a chance difference of just zero or one mutations per gene still has a cumulative probability of more than 5% under a model of equal underlying rates (Lynch and Katju 2004).

The Case for Subfunctionalization

Prior to the formal development of the DDC model, circumstantial evidence for duplicate-gene preservation via subfunctionalization of regulatory regions had been revealed through studies of polyploid fishes, which repeatedly demonstrated tissue specificity of expression of duplicated enzyme loci (Ferris and Whitt 1977, 1979). These observations have recently been supplemented by a substantial number of investigations in the zebrafish, a member of an ancient polyploid lineage that still retains roughly 25% of its original gene pairs in a functional state (Amores et al. 1998; Postlethwait et al. 2000). A key to such analyses has been the availability of orthologous single-copy genes in tetrapods. Comparison of zebrafish gene expression pat-

terns with those in the homologous tissues of tetrapod outgroup species (usually mouse or chicken) provides insight into the mechanisms by which zebrafish paralogs may have been preserved. In virtually every well-characterized case, subfunctionalization has been implicated.

Consider, for example, the two zebrafish genes for microphthalmia-associated transcription factor, *mitfa* and *mitfb*. Expression of the first gene is restricted to neural crest, and that of the second is restricted to the epiphysis and olfactory bulb (Lister et al. 2001; Altschmied et al. 2002). The products of these two zebrafish paralogs appear to be homologous to the two alternatively spliced forms of the product of the single-copy locus found in tetrapods, with subfunctionalization resulting from deletions in both regulatory and coding regions (such that each copy has adopted a single splicing variant). As another example, consider the two zebrafish cytochrome P450 aromatase genes (Chiang, Yan et al. 2001). One of these is expressed in the ovary and the other in the brain, whereas the orthologous single-copy gene in tetrapods is expressed in both tissues. Similarly, the zebrafish has two *sox11* genes, one of which is expressed in the anterior and the other in the posterior somites, whereas the single *sox11* product in the mouse is expressed in all somites (de Martino et al. 2000). Many other zebrafish genes appear to have become subfunctionalized following gene duplication (Westin and Lardelli 1997; Nornes et al. 1998; Force et al. 1999; Chiang, Pai et al. 2001; Quint et al. 2000; Bruce et al. 2001; McClintock et al. 2002; de Souza et al. 2005; Liu et al. 2005). Nevertheless, it should be noted here that the DDC model originally postulated subfunctionalization as a *process* of duplicate-gene preservation by degenerative mutation, whereas it is formally possible that *patterns* of partitioned gene subfunctions between extant duplicate genes may have arisen secondarily, with the initial phase of preservation having been driven by other mechanisms.

Observed patterns of subfunctionalization are by no means a peculiarity of polyploid fishes, and just a few additional examples are cited here. First, genome-wide analyses of the mRNAs of duplicate genes provide strong support for the idea that the partitioning of ancestral alternatively spliced variants, as seen in the case of the zebrafish *mitf* genes, is a common fate of duplicate genes throughout the animal kingdom (Kopelman et al. 2005; Su et al. 2006). Second, the nematode *Caenorhabditis elegans* has two β-catenin genes, one of which plays a role in cell signaling and the other in cell adhesion, whereas a single gene fulfills both functions in most other animals (Grimson et al. 2000; Korswagen et al. 2000). The functional differences between the paralogous *C. elegans* genes appear to be due to alterations in the coding region. Third, in the barnacle *Sacculina carcini*, two *engrailed* duplicates are expressed late in development. The expression of one member of this pair is restricted to the nervous system and that of the other to the epidermis, whereas a single gene is responsible for both expression patterns in other arthropods (Gibert et al. 2000). Fourth, all vertebrates harbor two members of the Snail developmental gene family, *Snail* and *Slug*, the summed expression of which is conserved among all species (Locascio et al. 2002).

Remarkably, however, the three major expression domains of the paralogs appear to have shifted from one copy to the other in various vertebrate lineages, presumably because of ongoing regulatory region exchange between the duplicates. Fifth, experiments in the mouse have shown that the combination of a coding region from a single member of a paralogous Hox gene pair with the regulatory elements of both pair members is sufficient for normal development (Tvrdik and Capecchi 2006). Although such manipulations may be viewed as effectively reversing the subfunctionalization process and recreating the ancestral single-copy locus, precise interpretations are rendered difficult by the large numbers of accumulated mutations in the paralogous copies. Finally, in maize (*Zea mays*), two copies of the *p1 myb*-like transcriptional activator partition expression patterns in male and female reproductive structures and leaves, whereas the single-copy ortholog in closely related teosinte is responsible for all of these expression patterns (P. Zhang et al. 2000). Numerous cases of suspected subfunctionalization in maize are associated with the reciprocal loss of conserved noncoding regions (presumably regulatory elements) (Langham et al. 2004).

These observations, and many more, show that there are numerous potential paths to subfunctionalization, including the reciprocal silencing of tissue-specific promoters, the adoption of alternative splicing forms, and modification of the coding regions of multifunctional proteins. In addition, although theoretical considerations have focused on the role of degenerative mutations in initiating the subfunctionalization pathway, other poorly understood processes may facilitate the process (Rodin and Riggs 2003). For example, in numerous animals, including nematodes and mammals, the X chromosome is inactivated in the male germ line. This means that any autosomal gene with male germ line expression will immediately lose that subfunction if duplicated to the X. The autosomal copy is then free to lose expression in somatic tissues, but forced to retain male germ line expression. Several autosomal/X pairs of *C. elegans* gene duplicates exhibit this pattern (Maciejowski et al. 2005). In addition, following allopolyploidization (the joining of genomes from two different species), plants often experience rapid rates of genomic rearrangement and/or abrupt epigenetic changes in methylation patterns, which can spontaneously induce restrictions in the tissue-specific expression patterns of paralogs (Adams et al. 2003; Osborn et al. 2003). Complementary reciprocal epigenetic silencing can lead to essentially instantaneous subfunctionalization, with the epigenetically silenced subfunctions eventually being replaced by neutrally accumulating background degenerative mutations, providing one mechanism by which subfunctionalization might proceed to completion even in very large populations.

Although unicellular species have no opportunities for tissue-specific specialization of gene expression, gene function partitioning may occur at the subcellular level (e.g., via the modification of transit signals for localization of proteins to specific organelles; Silva-Filho 2003; Schmidt et al. 2003). As discussed above, however, even if such physical opportunities exist, the probability of subfunctionalization is greatly reduced in popula-

tions of unicellular species with sufficiently large genetic effective sizes. Nonetheless, substantial evidence suggests that duplicate genes in the yeast *S. cerevisiae* are often victims of subfunctionalization. For example, empirical studies in which *S. cerevisiae* paralogs were swapped with the single-copy gene from the outgroup species *S. kluyveri* have provided compelling evidence for subfunctionalization at the level of coding DNA (van Hoof 2005). A gradual loss of ancestral subfunctions by paralogous copies of *S. cerevisiae* genes is suggested by observed reductions in the similarity of expression patterns (measured across different environments), the number of shared regulatory motifs, the number of shared interacting protein partners, and the number of shared functional domains with increasing age of paralogs (Gu et al. 2002b; Papp et al. 2003b; He and Zhang 2005b; van Hoof 2005). In contrast, the total numbers of regulatory motifs and of interacting protein partners for each member of a pair appear to remain roughly constant or even increase over time, suggesting an approximate balance between gains and losses of such elements (Papp et al. 2003b; He and Zhang 2005b), a pattern also found in mammalian duplicate gene pairs by Huminiecki and Wolfe (2004). The recruitment of novel regulatory elements is not surprising, as small patches of DNA are even more likely to be duplicated than entire genes. However, answers to the many questions raised by these observations will require additional work, like that of van Hoof (2005), with outgroup *Saccharomyces* species containing single-copy orthologs of the *S. cerevisiae* paralogs.

Speciation via the Divergent Resolution of Duplicate Genes

Most studies of duplicate genes have focused on their potential role in the origin of evolutionary novelties through the establishment of new gene functions. However, given the high rate at which duplicate genes arise, move to unlinked positions, and become randomly silenced or subfunctionalized, gene duplication may be an equally important contributor to the other major engine of evolution: the origin of new species (Lynch and Force 2000a; Shpak 2005). Consider an unlinked pair of duplicate autosomal genes in a diploid ancestral species, which then experience divergent silencing or subfunctionalization in two descendent lineages, effectively resulting in a map change (Figure 8.12). Because the F_1 hybrids of such lineages will be "presence/absence" heterozygotes at the two independently segregating loci, 1/4 of the F_1 gametes will contain null (absentee) alleles at both loci. Thus, if the gene is critical to gamete function, this single divergently resolved duplication will result in an expected 25% reduction in fertility. For a zygotically acting gene, 1/16 of the F_2 offspring from the interspecific cross will lack functional alleles at both loci, and another 1/4 will carry only a single functional allele. Thus, if the gene is haploinsufficient, 5/16 of the F_2 zygotes of such a cross will be inviable (and/or sterile). With n divergently resolved

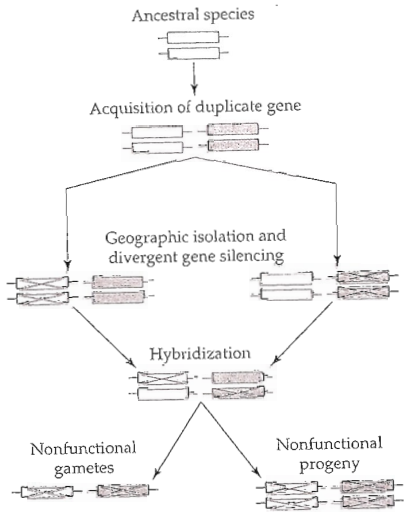

Figure 8.12 Divergent silencing of an ancestral duplicate gene in two geographically isolated lineages. One gene copy is denoted by a white box, the other by a green box, and an X denotes an inactivated gene. Gene pairs represent alleles at diploid loci. Progeny other than those indicated might have compromised fitness (e.g., individuals with just a single active gene).

duplicates, the expected fitness of hybrid progeny is $W = (1 - \delta)^n$, with δ denoting the reduction in hybrid fitness per map change. For example, with $\delta = 5/16$, $W = 0.024$ in the F_2 generation when $n = 10$, and $W = 5 \times 10^{-17}$ when $n = 100$. Observed rates of gene duplication indicate that this type of process is sufficiently powerful to yield nearly complete genomic incompatibility within a few million years of cessation of gene flow (Lynch and Force 2000a; Lynch 2002a; Shpak 2005), which is the approximate time scale over which postzygotic isolation generally occurs in animals (Parker et al. 1985; Coyne and Orr 1997; Sasa et al. 1998; Presgraves 2002; Price and Bouvier 2002).

Indirect evidence leaves little room for doubt that microchromosomal rearrangements resulting from gene duplication can be significant contributors to the establishment of postzygotic isolating barriers among species. Consider, for example, the mustards *Arabidopsis thaliana* and *Brassica oleraceae*, which are thought to have diverged from a common ancestor 10–20 MYA (Yang et al. 1999; Bancroft 2001). Although the two species contain many sets of orthologous chromosomal segments with long-range colinearity in gene content, alternative losses of orthologous gene copies in the two species are common (O'Neill and Bancroft 2000; Quiros et al. 2001). Complicating matters is the fact that *B. oleraceae* is a triploid derivative of the older lineage containing *A. thaliana*, which itself experienced two or three ancient polyploidization events, as described above. However, even this complexity is informative. For example, in an analysis of two triplicated paralogous

chromosomal regions in *B. oleraceae*, O'Neill and Bancroft (2000) found that about two-thirds of the paralogous groups had at least one member silenced on at least one paralog. These results, as well as others (e.g., Ku et al. 2000; Bennetzen and Ramakrishna 2002), suggest that the primary mechanism of chromosomal repatterning in plants may be duplication of large chromosomal regions followed by random loss of component genes. Estimated rates of microchromosomal rearrangement in plants range from 0.3 to 6.0 per lineage per million years (Lagercrantz 1998; Burke et al. 2003), and most of these estimates are downwardly biased by imperfect resolution in the associated mapping projects. Thus, considering that many rearranged fragments may contain multiple genes, the number of map changes per million years may commonly exceed ten in land plants.

Such high rates of microchromosomal rearrangement are by no means unique to plants. For example, comparing the genomic sequence of the nematode *Caenorhabditis elegans* with that of its congener *C. briggsae*, Coghlan and Wolfe (2002) estimated that 4,030 rearrangements had occurred over a period of 80 million years, implying 25 rearrangements per lineage per million years. Reciprocal exchanges between two genomic locations, local inversions, and transpositions from one location to another contribute to the total pool of rearrangements at a ratio of 1:1:2, and the vast majority of the segments involved span five or fewer genes. Rates of microchromosomal rearrangements in vertebrates appear to be at least as high as those in nematodes, with roughly 1.5 large-scale (> 100 kb) rearrangements occurring per lineage per million years, usually as a consequence of duplicative transposition, and smaller-scale rearrangements arising at rates of at least 20 per lineage per million years (McLysaght et al. 2000; Locke et al. 2003; Pevzner and Tesler 2003). In contrast, the rates of microchromosomal rearrangement in both *Drosophila* (fruit flies) and *Anopheles* (mosquitoes) appear to be substantially lower, in both cases about 7 per lineage per million years, with the vast majority of rearrangements being within, rather than between, chromosomes (González et al. 2002; Ranz et al. 2001, 2003; Sharakhov et al. 2002). Finally, genomic comparisons of the yeasts *S. cerevisiae* and *Candida albicans* imply a rate of microchromosomal rearrangement of about 2.3 per lineage per million years (Seoighe et al. 2000).

Although microchromosomal rearrangements need not always originate via gene duplication events, a large fraction of such events almost certainly do, for the simple reason that the movement of a gene to a new unlinked location is much more easily accomplished if the transition is made gradually while the original locus is still intact. The magnitude of gene duplication activity discussed above appears to be fully compatible with this interpretation. Under a steady-state birth/death process, a population is expected to gain $2NB$ new duplications per locus per generation, each with a probability of fixation equal to $1/(2N)$ under the assumption of initial neutrality. Assuming that the fate of most duplicates is nonfunctionalization, and recalling that each fixed duplication has a 50% probability of being resolved in favor of the new locus, the steady-state rate of origin of map changes is simply $(GB/2)$ per lineage per generation, where G is the total number of genes per genome. Using the estimates of B reported in Table 8.1 and the estimates of G from Table 3.2, the approximate expected number of map changes per million years is 2.5 for unicellular species, 15 for invertebrates, and 125 for vertebrates and plants. Not all such changes will necessarily induce hybrid incompatibilities, as some divergently resolved pairs of loci will be tandemly located and/or members of multigene families. Nevertheless, these collective results reinforce the idea that up to dozens of map changes per million years may passively accumulate in isolated lineages via small duplication events.

Duplication of autosomal genes is just one route by which map changes can be induced passively by divergent degenerative mutation, and to emphasize this point, we now consider three additional mechanisms. First, consider the situation in which an ancestral gene with a male-specific function is initially present on both sex chromosomes, with the copy on the X becoming silenced in one descendent lineage and the copy on the Y becoming silenced in a sister lineage. A cross between females of the first population (assumed to be XX) and males (XY) of the second would result in male progeny completely lacking in gene function, while the reciprocal cross would have active copies on both the X and the Y. Thus, duplication events involving genes on sex chromosomes have potential relevance to understanding the mechanisms underlying Haldane's rule, which states that interspecific genomic incompatibilities are most severe in heterogametic F_1 progeny (Orr 1997). A number of interesting reassignments of map locations of male-specific genes have been uncovered in mammals. For example, almost all extant mammals have autosomal *CDYL* and *CDYL2* genes, which carry out key housekeeping and testes-specific functions. However, in the lineage leading to humans, a copy of *CDYL* was duplicated to the Y chromosome, where it retained a function in spermatogenesis but lost the housekeeping function, while the autosomal loci experienced the opposite fate: loss of function in spermatogenesis but retention of the housekeeping role (Dorus et al. 2003).

Second, a remarkable set of examples of map changes induced by gene duplication in plants involves the movement of genes between organelle and nuclear genomes. Transfers of functional mitochondrial genes to the nuclear genome (accompanied by their subsequent loss from the mitochondrion) have occurred on many independent occasions within lineages of flowering plants, with the overall rate in some cases rivaling the rate of nucleotide substitution at silent sites (K. L. Adams et al. 2000, 2001, 2002). The details worked out for the mitochondrial respiratory protein gene *cox2* are particularly revealing. This gene was apparently duplicated to the nuclear genome of the ancestor of the Papilionoideae (a subfamily of legumes), transiently persisting as active copies in both genomes, with one or the other copy becoming randomly inactivated (in approximately equal frequencies, and by a variety of mechanisms) in almost all descendent lineages (Adams et al. 1999). Many closely related plant genera also exhibit

complex patterns of nuclear transfer of mitochondrial ribosomal protein genes (Adams et al. 2002). Moreover, these kinds of intergenomic gene transfers are not restricted to plant mitochondria, as a study of the chloroplast *infA* gene also reveals large numbers of transfers to the nuclear genome (Millen et al. 2001).

Third, although the previous arguments focus entirely on the divergent resolution of duplicate genes driven by degenerative mutation, map changes can also be induced by neofunctionalization, provided the copies acquiring new functions do so at the expense of the old function (Figure 8.13). Such changes, which arise whenever the ancestral locus takes on the new function while the descendent locus retains the original function, are expected to arise in 25% to 50% of cases of duplicate-gene neofunctionalization, depending on the population size (Lynch et al. 2001). Because the probability of preservation of duplicate genes by neofunctionalization increases with population size, unlike the many genetic theories of speciation that rely on population bottlenecks, this version of the gene duplication model is also effective in very large populations.

Genetic theories of speciation have traditionally focused on two competing sets of hypotheses, each of which has numerous adherents and detractors (reviewed in Orr 1996; Rieseberg 2001; Coyne and Orr 2004). The Dobzhansky–Muller model postulates the accumulation of lineage-specific gene sequence changes that are mutually incompatible when brought together in a hybrid genome, whereas the chromosomal model invokes the accumulation of rearrangements that result in mis-segregation in hybrid backgrounds. Both models are based on rather stringent assumptions. For example, the Dobzhansky–Muller model invokes the evolution of coadapted complexes of epistatically interacting factors, none of which have yet been identified at the molecular level, whereas chromosomal models generally focus on major rearrangements, the fixation of which can be greatly inhibited by the reduction in fitness in chromosomal heterozygotes.

A notable feature of the gene duplication model of speciation described above is that it is consistent with *both* the Dobzhansky–Muller model and the chromosomal model while requiring fewer assumptions than either of them. The gene duplication model is effectively a chromosomal model of speciation, but because the rearrangements are microchromosomal, and hence unlikely to cause significant problems during meiosis, they accumulate passively without any alteration in within-species fitness. The gene duplication model also masquerades as a Dobzhansky–Muller model in that the map changes induced by divergent resolution result in pseudo-epistatic interactions without any changes at the gene level. Genomic incompatibilities arising from reassignments of genes to new locations appear superficially as epistatic interactions because the loss-of-function phenotype is determined by the number of active alleles at the two homologous loci in hybrid progeny. Thus, low-resolution analyses of species incompatibilities that fail to identify the specific underlying loci can lead to misinterpretations regarding the underlying genetic mechanism of postzygotic isolation.

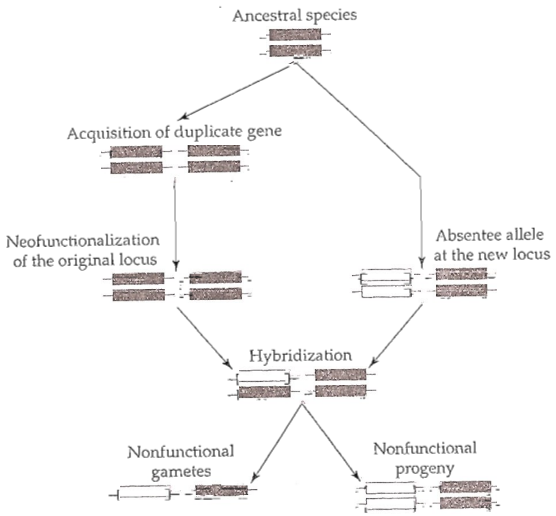

Figure 8.13 The origin of postzygotic incompatibility by neofunctionalization of an ancestral locus. A gene with the essential ancestral function is denoted by black, a neofunctionalized gene by red, and an empty locus by white. Neofunctionalization is assumed to have occurred at the expense of the original gene function. The species on the right either did not experience a duplication at the new locus or lost it prior to fixation.

The induction of map changes by neofunctionalization also blurs the distinction between species-isolating mechanisms based on adaptive genetic change and chromosomal rearrangements. For example, in a hybrid cross, a neofunctionalized ancestral locus would superficially appear to interact negatively with genetic factors residing in the single-copy species (see Figure 8.13). However, the incompatibility would not be a function of adaptive changes at the neofunctionalized locus, but simply an indirect consequence of the relocation of the ancestral gene function.

The vast majority of research on the genetic mechanisms of species-isolating barriers has focused on *Drosophila*, and most of this work, inspired by the classic Dobzhansky–Muller model, has concentrated on a search for "speciation" genes carrying the signature of local adaptation (Coyne and Orr 2004; Wu and Ting 2004). The genetic analysis of speciation is an extraordinarily difficult enterprise, as it is generally nearly impossible to determine whether a particular incompatibility factor played a formative role in species isolation or simply arose secondarily after interspecific gene flow had ceased.

Of even greater uncertainty, however, is whether a singular focus on adaptive mechanisms leads to a biased perspective on the origins of interspecific reproductive incompatibility. Given the generally high rate of microchromosomal rearrangement induced by gene duplication, an evaluation of the gene duplication model ought to be a priority of any study that seeks an unbiased perspective on the mechanisms of speciation, if for no other reason than to provide a null model against which hypotheses involving adaptation can be tested.

One remarkable observation that appears to be quite compatible with the gene duplication model of speciation involves the yeast *S. cerevisiae* and its close relatives, which exhibit numerous differences in gene order resulting from chromosomal rearrangements. Although the haploid offspring of crosses between such species are almost always sterile, after engineering the chromosomes to restore large-scale colinearity, Delneri et al. (2003) were able to increase fertility to 20%–30%. Some minor differences in gene order almost certainly went undetected in these constructs, so it is quite possible that restoration to complete colinearity would have had an even greater effect. Well-documented examples of the involvement of duplicate genes in the reproductive isolation of *Drosophila* species also exist. In two cases, a strong phase of positive selection operating on single paralogs has been implicated (Ting et al. 2004; Greenberg et al. 2006), but in some *D. melanogaster–D. simulans* hybrids, sterility appears to be a simple consequence of the movement of an essential gene to a new chromosomal location via an intermediate phase of gene duplication (and without a change in function), in full accordance with the model presented in Figure 8.12 (Masly et al. 2006). Other lines of evidence support the idea that isolating barriers are associated with small chromosomal rearrangements, although alternative mechanistic explanations (such as the capture of adaptive alleles by nonrecombining chromosomal inversions) have been suggested for such patterns (Noor et al. 2001; Rieseberg 2001; Navarro and Barton 2003).

Under the gene duplication model, certain groups of organisms are expected to be more prone to speciation than others, the most notable being lineages that experience a doubling in genome size. One potential example of such a key event was noted in Chapter 1: the colonization of ancestral eukaryotic genomes by endosymbiotic organelles. Considering the very large number of organelle-to-nucleus transfers that apparently occurred soon after the establishment of the mitochondrial progenitor (Martin et al. 1998), divergent resolution of duplicated organelle genes may have provoked the passive development of isolating barriers among a number of the basal eukaryotic lineages.

Polyploidization provides another enormous opportunity for the rapid proliferation of isolated lineages via the divergent resolution of duplicate genes. Following the first map changes induced by reciprocal silencing in sister polyploid taxa, the thousands of duplicate pairs still remaining in a functional state are free to become divergently resolved in subsequently isolated lineages, potentially yielding a large number of nested speciation

events. The origin of species via polyploidy-associated map changes may be especially common in plants, almost all of which are descendents of one or more polyploidization events (Werth and Windham 1991). In addition, particularly striking support for the gene duplication model of species isolation is revealed by the genomic structures of lineages derived from the yeast polyploidization event noted above (Scannell et al. 2006). Pairwise comparisons of the three completely sequenced yeast genomes (*S. castellii*, *S. cerevisiae*, and *Candida glabrata*) identify 100 to 200 divergently resolved paralogous gene pairs (approximately 5% of the single-copy genes in different species are not orthologs), and phylogenetic analysis places much of the genomic repatterning during the period of species emergence. Finally, although the details remain to be worked out, one of the most striking examples of cryptic speciation, the reproductive isolation of 14 morphologically indistinguishable members of the *Paramecium aurelia* complex, also appears to have developed shortly after an ancestral polyploidization event (Aury et al. 2006).

Key genome doubling events may have facilitated the diversification of many major animal lineages as well. First, given the apparent twofold-to-threefold increase in gene content in basal animals relative to fungi, genome amplification may have been involved in the origin of the major animal phyla. Second, as described above, it also appears that a substantial amount of gene duplication occurred prior to the radiation of the major vertebrate lineages. Finally, it may be no coincidence that the most species-rich lineage of vertebrates, the ray-finned fishes (~30,000 species; see Figure 8.5), is a descendent of an ancient polyploidization event. At least some evidence points to divergent resolution in isolated lineages in this group (Cresko et al. 2003), and secondary polyploidization events within specific lineages are associated with enhanced rates of speciation (Ferris et al. 1979; Taylor et al. 2001b; Hoegg et al. 2004; Mank and Avise 2006; Volff 2005). In contrast, all of the fish lineages that branched off prior to the basal polyploidization event (e.g., bichirs, bowfin, gars, paddlefish, and sturgeons) contain just a handful of species. These kinds of observations suggest that adaptive radiations are associated with polyploidization events not only because gene duplication opens up novel evolutionary pathways for the origins of new gene functions, but also because polyploidization generates a population genetic environment that is highly conducive to the passive origin of reproductive barriers.