CHAPTER 4

Gene duplication and evolution

Michael Lynch

Motivated in part by Ohno's (1970) influential book, substantial attention has been given to the idea that gene duplication is a major mechanism for the origin of new gene functions. A theoretical population-genetic framework for understanding the evolutionary mechanisms responsible for the success versus demise of gene duplicates has begun to emerge. Substantial evidence now exists that many of the key evolutionary lineages of multicellular eukaryotes have experienced one or more complete genome doublings (polypliodization) some time in the distant past (Wolfe 2001), and the newly unveiled genomic sequences of diverse species clearly indicate that gene duplication is an ongoing process in all organisms. The emerging picture is one in which the eukaryotic genome is a dynamic playing field in which new genes are continuously arising via duplication events, with most being eliminated by drift and/or natural selection, some simply replacing their ancestral copies, and a few being preserved along with their twins for long periods of time.

It is clear that the evolution of organismal complexity has been accompanied by a net growth in gene number. The genomes of the simplest single-celled prokaryotes appear to contain a minimum of 400 and a maximum of 7000 or so genes. The genomes of nonparasitic multicellular eukaryotes appear to contain no fewer than 10,000 genes, and the maximum probably exceeds 100,000. What remains unclear is whether the hallmarks of organismal complexity (e.g., the origin of new cell types, mechanisms of cell–cell communication, cell adhesion, etc.) require an amplification of genome size. There are, after all, many ways to wring multiple functions out of single-gene copies, including alternative splicing, modularization of tissue-specific expression patterns, post-translational modification, etc.

This is not to say that gene duplication is a minor player in the origin of new gene functions. Numerous compelling cases for neofunctionalization following duplication events have been lucidly outlined by Pithy (1999). But the broader view taken in this chapter is that gene duplication influences evolution via processes other than the evolution of new functions. Most notable among these processes are: (1) the elimination of pleiotropic constraints as duplicate descendants partition up the multiple functions of their ancestral single-copy gene, and (2) the passive origin of microchromosomal rearrangements when ancestral gene functions are reassigned to new locations. In this sense, the gene-duplication process provides fuel for both of the major engines of evolution—adaptation and 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, the complete genomic sequences for several species provide an indirect route (Lynch and Conery 2000, 2003). Through comparative sequence analysis, the total pool of duplicate genes within a genome can be identified, and the relative ages of the duplicate pairs can be estimated from the pairwise divergence of silent sites in coding regions under the assumption that such sites accumulate nucleotide changes at a relatively constant rate. The age distribution of duplicate pairs can then be used
to estimate the average rates of origin and elimination of duplicate genes, in the same manner that demographers use age distributions to estimate birth and death rates. Under a steady-state birth–death process, advocated in a somewhat different context by Nei et al. (1997, 2000), the frequency of duplicate pairs will exhibit an exponential decline with age, with the abundance of identical to nearly identical duplicates providing information on the birth rate, and the rate of decline in abundance with age providing an estimate of the loss rate. Most eukaryotic genomes assayed to date exhibit the approximate pattern predicted by this model (Lynch and Conery 2000, 2001, 2003; Achaz et al. 2001).

Application of the preceding logic to the genomic sequences of a diverse set of eukaryotic species (Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Encephalitozoon cuniculi) yields estimates of the rate of birth of duplicate genes in the range of 0.001–0.03 per gene per million years, with an average of or 0.01 per gene per million years (Lynch and Conery 2003). In other words, on a time scale of 100 million years or so, nearly every gene in a genome can be expected to have duplicated at least once. These indirect estimates are conservative in that they do not include contributions from large multi-gene families, of which there are many in eukaryotes. In addition, direct empirical estimates of gene duplication rates in Drosophila appear to be higher, on the order of $10^{-9}$ to $10^{-11}$ per gene per generation (Shapiro and Finnerty 1996). Thus, the rate of duplication per gene is at least of the same order of magnitude as the rate of mutation per nucleotide site (Li 1997), and perhaps considerably higher. This implies that changes in gene content may often rival changes in gene sequence as a mechanism of phenotypic evolution. On the other hand, as is the case for most replacement nucleotide substitutions within genes, most duplicate genes appear to be evolutionarily short-lived. The half-lives for such genes in the previous list of species range from 1 to 17 million years, with an average of about 5 million years (Lynch and Conery 2003).

Although the individual estimates are very approximate, these demographic analyses highlight the dynamic nature of eukaryotic genomes with respect to gene content. As a consequence of a stochastic balance between gene birth and death rates, total genome size may remain approximately constant within specific lineages for long periods of time, but throughout such periods there is likely to be continual turnover with respect to the specific genes that are present in redundant copies. The precise mechanisms by which duplicate genes arise are not yet well understood, but they are probably diverse. Many newly arisen gene duplicates apparently arise via local events, as they are often tandemly associated with their parental copy. However, duplications to new chromosomal locations also occur. The most common mechanism for nontandem duplicates may be the capture of nascent DNA fragments during the repair of double-strand breaks, which arise several times daily per cell (Ricchetti et al. 1999; Lin and Waldman 2001), although the source of such DNA fragments is uncertain. Whether duplication spans will contain one or more fully functional genes is a matter of chance, and incomplete duplication events may result in products that are “dead on arrival.”

**Mechanisms for the preservation of duplicate genes**

All duplicate genes are expected to be initially carried by a single member of a population, and hence to be highly vulnerable to stochastic loss early in their history. To be successful in the long term, a duplicate gene must first drift towards fixation, and then having arisen to high frequency, the selective forces for its maintenance must be sufficiently large to prevent its subsequent loss by degenerative mutation. The precise mechanisms by which duplicate genes are preserved 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 genome size but does induce an alteration of the genetic map (the relevance of which is discussed below).

**Neofunctionalization**

One of the more notable mechanisms for the joint preservation of a pair of gene duplicates is the
process of neofunctionalization, whereby one copy acquires a beneficial mutation to 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 essential wild-type function should have had no barriers to fixation prior to duplication. Gene duplication alters the selective landscape for the subset of beneficial mutant alleles with negative pleiotropic effects on the ancestral gene function by opening up the opportunity for one locus to experiment evolutionarily while the other retains the ancestral function. This model extends back at least to Haldane (1933), and some aspects of it were explored quantitatively by Walsh (1995) for the case in which the duplication is assumed to be initially fixed in the population.

Although most considerations of the neofunctionalization process have focused upon mutations arising subsequent to the duplication event, Spofford (1969) made the key observation that the arrival of new mutations may not be a prerequisite for neofunctionalization. The simple logic for this argument is that the spectrum of mutations arising subsequent to a duplication event are the same as those arising prior to duplication. Thus, a mutant allele endowed with a beneficial function at the expense of an essential ancestral function may be maintained at low frequency at the ancestral single-copy locus by balancing selection—even though homozygotes for such an allele will be inviable, the heterozygotes may have elevated fitness.

The presence of such ancestral polymorphisms can facilitate the route to neofunctionalization in two ways. First, if the duplicate locus is founded by a neofunctional allele, fixation at the new locus will be promoted by positive selection while the original locus retains the ancestral function. Alternatively, if the new locus is founded by a “wild-type” allele that rises to a high enough frequency, the selective regime at the ancestral locus will be altered to one of positive selection for fixation of the neofunctional allele. In either case, the final outcome is functionally equivalent to the fixation of an overdominant gene action, with one locus being essentially monomorphic for the “wild-type” allele and the other for the neofunctional allele.

The primary conditions necessary for the maintenance of neofunctional alleles at the ancestral locus have been worked out (Lynch et al. 2001). First, the mutation rate to null alleles must be less than the square of the selective advantage of the neofunctional allele in the heterozygous condition ($s^2$). Second, the effective population size ($N$) must be sufficiently large that the power of random genetic drift is less than the strength of balancing selection (approximately, $N > 4/s^2$). Provided these conditions are met, then the neofunctional allele will be present at the initial locus with approximate frequency $s$, for example, an allele that is lethal in the homozygous state but increases fitness by 5 percent in the heterozygous state would have an expected frequency of ca 5 percent. Thus, just a moderate heterozygous selective advantage combined with a moderately large effective population size provides a setting by which a population can be poised to proceed towards neofunctionalization following a duplication event.

The actual probability of preservation of a pair of duplicate genes by neofunctionalization depends on several additional factors (Lynch et al. 2001). First, 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 if the new locus is founded by a wild-type allele, that allele will have a selective advantage that depends on the frequency of neofunctional homozygotes at the ancestral locus ($= s^2$). This advantage occurs because “absentee” homozygotes at the new locus are lethal on this genetic background. Alternatively, if the new locus is founded by a neofunctional allele, the marginal selective advantage of the founder allele over the absentee allele is $= s^2$ (Lynch et al. 2001). In both cases, the probability of neofunctionalization scales with $s^2$ because the probability of fixation of a newly arisen mutation in large populations is equal to twice the selective advantage (Crow and Kimura 1970). A second factor that influences the probability of neofunctionalization is the degree to which the two loci are linked. In the case of complete linkage (as is approximately the case for a pair of tandem duplicates), neofunctional alleles
The masking effect of duplicate genes

Because all loci harbor suboptimal alleles due to the recurrent introduction of deleterious mutations, one might imagine that duplicate genes have an intrinsic selective advantage associated with their ability to mask the effects of recessive deleterious mutations at the ancestral locus. However, the magnitude of such an indirect advantage may only rarely be strong enough to promote the permanent preservation of duplicate genes. Fisher (1935) realized that two genes with identical roles in an effectively infinite population will not be mutually maintained by selection unless their mutation rates to nulls 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 fact, because random genetic drift will eventually lead to the stochastic loss of one locus in a finite population, not even identical mutation rates will be sufficient for the permanent retention of duplicate genes (Clark 1994; Lynch et al. 2001). Consider, for example, the masking of a deleterious recessive lethal. The equilibrium frequency of null homozygotes at a single-copy locus is equal to the null mutation rate $\mu$ in large populations (Crow and Kimura 1970), and this must also equal the selective advantage of a rare functional duplicate at a new locus. However, $\mu$ is also the rate of silencing of the new allele, so these two factors cancel exactly, rendering the new duplicate effectively neutral. Thus, in the absence of neofunctionalization, something additional is required for the maintenance of a duplicate gene.

One possibility is a direct selective advantage of a duplicate locus, that is, increased fitness in individuals with three versus two functional genes. However, this scenario must be tempered with the alternative possibility that the overexpression of duplicate loci may negatively influence the quantitative balance between the total expression of the pair and their interacting partners (except in the case of polyploidization events, which maintain the stoichiometric relationships among all pairs of genes). This different outcome of incremental duplications and whole-genome duplications may explain the apparently higher rates of retention of
duplicate genes following polyploidization events (Lynch and Conery 2000). A second possible mechanism for the permanent retention of duplicate genes is a balance between the differential selective advantages and mutation rates of a pair of duplicate loci. This would require, for example, one gene to operate more efficiently and the other to have a lower mutation rate to nulls (Nowak et al. 1997). Whether the fine balance required under this scenario is likely to ever be met is unclear. Still another possibility is the maintenance of a duplicate gene by recurrent introduction of new copies by the duplication process itself. In this case, however, since independent introductions of new copies are expected to almost always appear in new genomic locations, this process would at best lead to an equilibrium level of overall genomic redundancy, but not to the permanent retention of any particular pair of loci (Wagner 2000a). In other words, although the total amount of redundancy in a gene family may remain relatively constant over time under this model, the specific members of the family and their genomic locations would be expected to turn over as a consequence of the steady-state birth–death process, presumably leading to L-shaped age distributions similar to those actually observed (Lynch and Conery 2000).

Although the masking effects described above are concerned with compensation for mutationally silenced alleles, it has also been argued that duplicate genes may provide a buffer against cellular mishaps that lead to localized absence of gene expression (from normally active genes) (Tautz 1992; Nowak et al. 1997). There are a number of potential sources for such developmental errors, including somatic mutations, errors in transcription and translation, and errors in the inheritance of methylation patterns. Following the logic outlined above, for this mechanism to maintain a duplicate gene by natural selection, the developmental error rate would have to exceed the rate of origin of null mutations at the duplicate locus.

In summary, although the various masking models for the preservation of duplicate genes cannot be formally rejected, they all require rather special sets of mutational conditions and enormous population sizes to enable the very weak selective advantages of redundancy to come to prominence. The general paucity of duplicate genes in haploid microbes raises the most serious challenge to the idea that masking plays a prominent role in duplicate-gene retention.

Subfunctionalization

Given the limitations on the various masking hypotheses, neofunctionalization is often assumed to be the only mechanism by which duplicate genes can become permanently preserved. Under this assumption, because neofunctionalizing mutations are rare relative to degenerative mutations, the vast majority of new gene duplicates are expected to be lost within a relatively short period of time. 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 \( 2 \ln(2N) \) generations (Kimura and Ohta 1969), and for the small remaining fraction \( 1/(2N) \) that manage to drift to fixation, subsequent silencing of one copy by degenerative mutations will eventually ensue. Most of the theory on this matter has assumed a model for fitness in which all genotypes are equally viable except for the lethal double-null homozygotes. In this case, provided \( N \mu_s << 1 \), the average time to gene silencing is approximately equal to the mean waiting time until the appearance of a null mutation at one of the loci, \( 1/(2\mu_s) \) generations. On the other hand, for \( N \mu_s >> 1 \), the time to silencing is no longer limited by the mutation process, but is on the order of the time for an effectively neutral null mutation to drift to fixation, that is, \( 4N \) generations (Watterson 1983; Lynch and Force 2000a). These predictions of a relatively rapid demise of the vast majority of duplicate genes are inconsistent with the high levels of duplicate-gene retention observed in ancient polyploid lineages (Wagner 1998; Force et al. 1999). Since the evidence for the origin of new gene functions in these lineages is limited, alternative mechanisms for duplicate-gene preservation must be at work.

A potentially powerful mechanism becomes apparent when one considers a broader view of gene structure than assumed under the classical model. Because of the complex nature of regulatory
regions, most genes in multicellular eukaryotes have independently mutable subfunctions. That is, a mutation that causes the loss of gene expression in one particular tissue or developmental period does not necessarily affect other tissue- or tissuespecific aspects of expression. This more general view of gene expression leads to the prediction that duplicate-gene preservation must sometimes result from the partitioning of ancestral gene functions through complementary loss-of-function mutations in paralogous copies (Fig. 4.1). Under the duplication—degeneration—complementation (DDC) model of Force et al. (1999), this subfunctionalization process is driven entirely by degenerative mutations, which we know to be much more common than beneficial mutations.

Duplicate-gene preservation by subfunctionalization is a two-step process. First, one of the genes must become fixed for a mutation that eliminates a particular subfunction, an event that permanently preserves the second copy. Then, the second copy must lose an alternative subfunction, thereby reciprocally preserving the first copy. If the effective population size is sufficiently small that segregating null mutations are typically rare ($N_{we} \ll 1$), then the fate of a newly arisen duplicate gene under the DDC model depends almost entirely on the relative rates of origin of subfunctionalizing and nonfunctionalizing mutations ($\mu_\ast$ versus $\mu$), and the probability of subfunctionalization can be approximated with combinatorial logic.

Consider, for example, a gene with two independently mutable subfunctions. The probability that a newly arisen duplicate will drift to fixation is $1/(2N)$, and having arrived at this point, the probability that the first fixed mutation eliminates a subfunction from one of the genes is simply the fraction of mutations that are of the subfunctionalizing type $2\mu_\ast/(2\mu_\ast + \mu)$. After such a fixation event, the intact locus is no longer free to lose its now unique subfunction whereas the partially debilitated locus is free to become completely silenced, so the total permissible mutation rate for the next step is $(2\mu_\ast + \mu)$, with $\mu_\ast/(2\mu_\ast + \mu)$ being the fraction of permissible mutations that eliminate the complementary subfunction at the intact locus. The probability of
subfunctionalization ($P_{sub}$) is equal to the product of these three probabilities, $\alpha^2/(4N)$, where $\alpha = 2\mu_s/(2\mu_s + \mu_r)$ is the fraction of degenerative mutations that eliminate a single subfunction. Under this two-subfunction model, $P_{sub}$ approaches a maximum of $1/(4N)$ as $\alpha \to 1$.

A number of factors can increase the probability of subfunctionalization above this level, but none of them increases the upper bound to a level greater than $1/(2N)$. This is to be expected since $1/(2N)$ is the probability of initial fixation of an entirely neutral duplication, and the subsequent probability of permanent preservation cannot exceed one. For example, increasing the number of independently mutable subfunctions ($z$) increases $P_{sub}$ by increasing the number of paths by which complementary loss-of-function mutations can be acquired by the two copies (Force et al. 1999). Generalizing the definition of $\alpha$ to $z\mu_s/(z\mu_s + \mu_r)$, and keeping the ratio of subfunctionalizing to nonfunctionalizing mutation rates ($z\mu_s/\mu_r$) constant while allowing $z \to \infty$, it can be shown that the upper limit to $P_{sub}$ is $\alpha^2/(2N)$ (Lynch and Force 2000a). Thus, for the extreme case in which all mutations eliminate only single subfunctions and there are an effectively infinite number of such subfunctions, $P_{sub}$ attains a maximum value of $1/(2N)$.

Although the preceding theoretical results apply to the situation in which all mutations completely eliminate one or all subfunctions of a gene, mutations with partial effects on gene expression will further increase the probability of duplicate-gene preservation, even providing a retention mechanism for duplicate genes whose expression patterns cannot be subdivided (Lynch and Force 2000a). Such preservation occurs whenever the functional 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. Consider, for example, a gene with a single function, with $s$ being the number of mutations with partial effects necessary to completely eliminate gene function. Letting $\mu_s$ be the rate of origin of such mutations and $\rho = \mu_s/(\mu_s + \mu_r)$ be 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., $s \to \infty$). Again, this limit approaches a maximum value of $1/(2N)$ as mutations with partial effects become more predominant (i.e., as $\rho \to \infty$).

Averof et al. (1996) suggest still another mechanism by which the preservation of duplicate genes by subfunctionalization can be promoted. Newly arisen duplicates need not always be complete since, for example, critical regulatory elements in the flanking regions may be missing at the time of the duplication event. Recall that the probability of subfunctionalization of a newly arisen duplicate with two intact subfunctions is $\alpha^2/(4N)$ under a model in which single mutations completely eliminate aspects of gene expression. If a gene duplicate is missing one subfunction at birth, then the first step towards subfunctionalization has already been met, and the probability of subfunctionalization increases to $\alpha/(4N)$.

Finally, we note that the preceding theory applies to the case in which the effective size of a population is small (specifically, $N\mu_r << 1$). Provided these conditions are met, the simultaneous presence of polymorphisms at both loci is rare, and the probability of subfunctionalization is essentially independent of both the population size and the degree of linkage between duplicates. For larger populations, however, linkage plays a key role in determining the probability of subfunctionalization (Lynch et al. 2001). For unlinked duplicates, $P_{sub}$ asymptotically approaches zero at large $N$. Such behavior is a consequence of the long time (approximately $4N$ generations) required for a newly arisen duplicate to drift to initial fixation. If $N$ is sufficiently large, essentially all descendants of the initial duplicate will acquire silencing mutations by the time the lineage becomes fixed. On the other hand, for completely linked duplicates, $P_{sub}$ increases with increasing $N$, asymptotically approaching $1/(2N)$ as $N\mu_r \to \infty$. This behavior results from the fact that a linked pair of duplicates has a weak selective advantage over a single-copy gene, because complete inactivation of a "two-copy" allele requires the silencing of both members of the pair.

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. However, this preservational process 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 is a master-of-none” syndrome, such that an adaptive conflict exists between its multiple subfunctions. Under these conditions, complementary loss-of-subfunction mutations are expected to alter the selective landscape experienced by the two members of a duplicate pair, enabling each copy to become more refined to its specific subset of tasks (Piatigorsky and Wistow 1991; Hughes 1994). By this means, gene duplication combined with degenerative mutations may provide a unique mechanism for the creation of novel evolutionary opportunities through the elimination of pleiotropic constraints. Thus, although neofunctionalization and subfunctionalization may be viewed as independent preservational mechanisms, one involving positive selection for new beneficial functions and the other involving only the chance fixation of degenerative mutations, the processes of initial preservation and subsequent modification may become mutually blurred late in the evolutionary history of a pair of gene duplicates.

The case for subfunctionalization

Prior to the formal development of the DDC model, circumstantial evidence for subfunctionalization as a mechanism for duplicate-gene preservation had already accumulated through a series of isozyme studies in polyploid fishes that repeatedly demonstrated the presence of tissue-specificity of expression of duplicated enzyme loci (Ferris and Whitt 1977, 1979). These observations have recently been supplemented by a substantial number of DNA-based investigations in zebrafish. Genomic analysis has shown the zebrafish to be a member of an ancient polyploid lineage (Amores et al. 1998), and on the order of 25 percent of the thousands of original pairs of duplicates arising from the polyploidyization event are still functional (Postlethwait et al. 2000). A key to understanding the evolutionary modifications of these surviving zebrafish duplicates has been the availability of orthologous single-copy genes in tetrapods (usually chicken or mouse) as outgroups. Comparison of gene-expression patterns in the homologous tissues of these species provides a means of evaluating whether the differences among zebrafish paralogs is a result of the partitioning of ancestral functions or of the origin of new functions. In virtually every well-characterized case, subfunctionalization appears to be the most likely mechanism of preservation.

Consider, for example, the two zebrafish genes for microphthalmia-associated transcription factor, mitfa and mitfb (Lister et al. 2001). Only mitfa is expressed in neural crest, and only mitfb is expressed in the epiphysis and olfactory bulb. Although some of the molecular details remain to be worked out, the two zebrafish genes appear to be homologous to the two alternatively spliced forms of the single-copy locus found in tetrapods, with subfunctionalization (each copy adopting a single splicing variant) resulting from deletions in both regulatory and coding regions. As another example, consider the two zebrafish cytochrome P450 aromatase genes (Chiang et al. 2001b). One of these is expressed in the ovary and the other in the brain, whereas the single-copy gene in tetrapods is expressed in both tissues. Likewise, zebrafish has two soxII genes that are orthologous to the single-copy gene found in tetrapods. Although the two zebrafish paralogs overlap considerably in expression pattern, soxIIa is expressed in the anterior and soxIIb in the posterior somites, whereas the single soxII product in mouse is expressed in all somites (de Martino et al. 2000). Other zebrafish genes that appear to have experienced subfunctionalization following gene duplication include dlx (Quint et al. 2000), en1 (Force et al. 1999), hoxb5 (Bruce et al. 2001), notch (Westin and Lardelli 1997), pax6 (Nornes et al. 1998), and sox9 (Chiang et al. 2001a).

The process of subfunctionalization is by no means a peculiarity of polyploid fishes. For example, the nematode Caenorhabditis elegans has two β-catenin genes, one playing a role in cell signaling and the other in cell adhesion, whereas a single gene fulfills both functions in most other metazoa (Grimson et al. 2000; Korswagen et al.
The functional differences between the paralogous *C. elegans* genes appear to be due to alterations in the coding region. In the barnacle *Saccocirina carinii*, two engrailed duplicates are expressed late in development, one restricted to the nervous system and the other to the epidermis, whereas both expression patterns are fulfilled by a single gene in other arthropods (Gibert et al. 2000). In maize (*Zea mays*), two copies of the *p1 myb*-like transcriptional activator partition up expression patterns in male and female reproductive structures and leaves, whereas the single-copy orthologue in closely related teosinte fulfills all of these expression patterns (Zhang et al. 2000).

In addition to being facilitated by mutations in regulatory regions, subfunctionalization can occur via changes in coding regions, either through substantial modifications of alternative domains in proteins with multiple functions (qualitative subfunctionalization, as in the β-catenin genes of *C. elegans*) or through the accumulation of mildly deleterious mutations in both copies of a single-function gene (quantitative subfunctionalization). Insight into the types of selective pressures operating on the coding regions of duplicate genes may be acquired by comparing the number of nucleotide changes per amino acid replacement site (R) and per silent site (S). An R/S ratio greater than one implies directional selection for change, as expected under a scenario in which one or both members of a pair have evolved a new function, whereas a ratio less than one implies that selection is predominantly purifying in nature, and a ratio not significantly different from one is consistent with an absence of selection. Large-scale surveys of duplicate genes arising from ancient polyploidization events in mammals (Li 1985), *Xenopus laevis* (Hughes and Hughes 1993), and zebrafish (Van de Peer et al. 2001) have consistently found R to be substantially less than S, with the average ratio being similar to that observed in interspecific comparisons of single-copy genes.

One limitation of these types of comparative analyses is the cumulative nature of the nucleotide differences, which represent the culmination of tens of millions of years of evolution. In principle, newly arisen gene duplicates might experience an early phase of directional selection (with one member evolving a new function) or of relaxed selection (as a consequence of functional redundancy) followed by a longer phase of purifying selection once new gene functions have become established. A comparative analysis of distantly related sequences would reveal only the average of these patterns. To evaluate whether duplicate genes experience changes in intensities of selection as they age, Lynch and Conery (2000, 2003) estimated R and S for the complete sets of gene duplicates in fully to partially sequenced genomes of several plants, animals, and fungi. For all species, the R/S ratio declines with S. Although a considerable fraction of newborn duplicates (those with S ~ 0) appear to accumulate replacement substitutions at rates not significantly different from the neutral expectation, and a few even exhibit the signature of directional selection (R/S > 1), the average R/S for newborn duplicates in eukaryotes is approximately 0.7. This declines by about tenfold as S increases, with the vast majority of pairs appearing to experience strong purifying selection by the time they have diverged by 10 percent at silent sites. Thus, the average intensity of purifying selection operating on a pair of duplicate genes does indeed increase with the age of the pair.

By comparing the coding regions of two duplicate genes to that of a single-copy outgroup, one may inquire further as to whether the R/S ratio differs among the duplicate copies. Using this approach, Van de Peer et al. (2001) found that about half of the pairs of paralogous zebrafish genes exhibit significant heterogeneity in R/S between the two members of the pair. Why the difference? An intriguing possibility concerns the observation that the rate of replacement substitution in vertebrate genes is inversely correlated with the number of tissues in which the gene is expressed (Hastings 1996; Duret and Mouchiroud 2000), that is, genes with more restricted tissue-specific patterns of expression evolve more rapidly at the amino acid level. Since a large fraction of zebrafish paralogues appear to have undergone subfunctionalization, it will be of considerable interest to learn whether the more rapidly evolving members of pairs are also the ones that have retained fewer functions. The one locus for which data are available, triosephosphate isomerase, follows the expected pattern (Merritt and Quattro 2001).
The analyses of genewide substitution rates noted above potentially hide a significant amount of information on protein divergence. For example, it is possible for the two members of a duplicate pair to evolve at the same average rate, with each member experiencing a different spatial pattern of evolutionary change, for example, one having a higher rate at the 3' end and the other a higher rate at the 5' end. To evaluate this possibility, Dermizakis and Clark (2001) introduced a computational method for testing for spatial variation in the substitution pattern in the coding regions of different paralogues. Approximately 50 percent of their comparisons of mouse and human duplicate genes revealed significant regional variation among paralogous copies, some of which appear to be associated with functional domains.

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, the high rate at which duplicate genes arise, move to unlinked positions, and become randomly silenced or subfunctionalized suggests that gene duplication may be an equally important contributor to the other major engine of evolution—the origin of new species (Lynch and Force 2000b). Consider an unlinked pair of duplicate autosomal genes in a diploid ancestral species. Divergent silencing or subfunctionalization of the duplicate copies in two descendents species results in a map change that can secondarily induce incompatibilities in hybrid progeny in several ways. Because the F1 hybrids will be “presence-absence” heterozygotes at the two independently segregating loci, one-fourth of the F1 gametes will contain null (or absentee) alleles at both loci. Thus, for a gene that is critical to gamete function, this single divergently resolved duplication would result in an expected 25 percent reduction in fertility. For a zygotically acting gene, 1/16 of the F1 offspring from the interspecific cross would lack functional alleles at both loci, and another one-fourth would carry only a single functional allele. Thus, if the gene is haploinsufficient, 5/16 of the F1 zygotes of such a cross would be inviable (or sterile).

Indirect evidence leaves little room for doubt that microchromosomal rearrangements resulting from gene duplication are significant contributors to the establishment of postzygotic isolating barriers among species (Lynch and Force 2000b), and direct support is beginning to emerge from genomic analyses (Lynch 2002). For example, comparative-mapping data from plants have identified pervasive microchromosomal rearrangements associated with duplication events (see Bancroft 2001 for a review). Consider the mustards Arabidopsis thaliana and Brassica oleracea, which are thought to have diverged from a common ancestor 10–20 million years ago (Yang et al. 1999). The two species exhibit many sets of paralogous chromosomal segments with substantial long-range colinearity in gene content, but small-scale lineage-specific rearrangements resulting from alternative losses of orthologous gene copies are common (O’Neill and Bancroft 2000; Quiros et al. 2001). Complicating matters is the fact that B. oleracea is a triploid derivative of the older lineage containing A. thaliana, which itself experienced at least one ancient polyploidization event (prior to the divergence of Arabidopsis and Brassica). However, even this complexity is informative. For example, in an analysis of two triplicated paralogous chromosomal regions in B. oleracea, O’Neill and Bancroft (2000) found that about two-thirds of the component genes had at least one member silenced on at least one paralogue (Fig. 4.2). These results as well as others (e.g. Ku et al. 2000) suggest that the dominant mechanism of chromosomal repatterning in plants is duplication of chromosomal regions followed by random gene loss. Although the data are less extensive for animals, recent chromosomal comparisons in mammals (Dehal et al. 2001) and nematodes (Coghlan and Wolfe 2002) indicate that plants are by no means unique in this regard.

To emphasize that the inheritance of ancestral duplications of autosomal genes is just one route by which reproductive incompatibilities can passively arise from map changes induced by the divergent resolution of duplicate genes, we now consider four additional mechanisms. First, consider the situation
in which an ancestral duplicated gene for 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 being silenced in a sister lineage. A cross between females of the first population and males of the second would result in male progeny completely lacking in function, while the reciprocal cross would have active copies on both the X and the Y, potentially leading to dosage problems.

Other complexities can arise with duplicated genes distributed on an autosome and a sex chromosome, particularly if gametic imprinting occurs (Lynch and Force 2000b). These types of imbalances are not expected in F₁ females (assuming the male is the heterogametic sex) because female progeny inherit identical genomic complements from both parents. Thus, duplication events involving genes on sex chromosomes are of potential relevance to understanding the mechanisms underlying Haldane's rule, which states that incompatibilities in interspecific crosses are most severe in progeny of the heterogametic sex (Orr 1997).

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. RNA-mediated transfers of several mitochondrial genes to the nuclear genome (accompanied by subsequent loss from the mitochondrion) have occurred on many independent occasions within recent lineages of flowering plants, with the overall rate in some cases rivaling the rate of nucleotide substitutions at silent sites (Adams et al. 2000, 2001). 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 (with approximately equal frequencies, and by a variety of mechanisms) in almost all descendent lineages (Adams et al. 1999). These types of intergenomic gene transfers, initiated by gene duplication events, are not restricted to the mitochondrial genes, as a study of the chloroplast infA gene indicates large numbers of transfers to the nuclear genome (Millen et al. 2001).

What are the implications of such organelle-to-nucleus transfers? Consider the cross between a female of a species with an autosomal copy (A) of a gene (but no organelle copy, m) and a male of another species with the reciprocal arrangement. Letting small letters denote absentee alleles and assuming maternal inheritance of the organelle, the F₁ cross yields Aa/m progeny. If the gene is active in the gametic state, these individuals are expected
to experience a 50 percent reduction in fertility due to the production of null (a/m) gametes. For a zygotically active gene, ignoring potential problems of gene dosage, there will be a 25 percent loss of progeny in the F$_2$ generation due to the segregation of aa/m genotypes. With $n$ such organelle-gene relocations, the fraction of viable F$_2$ progeny will be reduced to $(3/4)^n$. Thus, when one considers the very large number of organelle-to-nucleus transfers that apparently occurred soon after the colonization of early eukaryotes by the progenitors of mitochondria and chloroplasts, it is quite conceivable that divergent resolution of duplicated organelle genes played a significant role in the development of complete isolating barriers among the major eukaryotic lineages.

Third, although the previous arguments focus entirely on driving the divergent resolution of duplicate genes 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. As noted above, the probability of preservation of duplicate genes by neofunctionalization increases with population size because of the increased number of targets for rare beneficial mutations. Thus, unlike many genetic theories of speciation, the gene-duplication model works in very large populations. Moreover, contrary to the assumption of models that associate speciation with presumptive incompatibilities of independent adaptive changes acquired by sister taxa, under the gene-duplication model for speciation, neofunctionalization induces incompatibilities entirely through change in the map position of the ancestral gene function, that is, no complex mechanisms of epistasis associated with “coadaptive gene complexes” are involved.

Fourth, duplicate genes arising subsequent to prezygotic isolation of lineages are highly relevant to the process of divergent resolution. As noted above, duplication events are ongoing processes in all lineages, and because they are distributed over a large number of loci, those experienced in different lineages will almost certainly involve unique genes. To gain an appreciation for the rapidity with which map-change induced genomic incompatibilities can arise via ancestral and recurrent gene-duplication events, we now consider a simple model for genes with a single function. Letting $b$ denote the rate of origin of new gene duplicates and $d$ denote the rate of silencing, the dynamics for the number of excess copies of a functional locus are given by

$$\frac{dn}{dt} = b(n + 1) - dn, \quad (4.1)$$

which yields the equilibrium number of excess copies of a gene, $b/(d - b)$. From the empirical results presented above, $b$ is on the order of $10^{-8}$ to $10^{-7}$ per generation, and for the situation in which all gene duplicates are subject to eventual silencing, $d$ is equivalent to the rate of origin of silencing mutations ($\sim 10^{-3}$ per generation). Thus, we expect on the order of 0.01–1 percent of the loci in a typical eukaryotic genome to be in a duplicated state at any point in time. These levels are consistent with the observed numbers of young duplicates in full-genomic sequences (Lynch and Conery 2000). From Watterson (1983) and Lynch and Force (2000a), the mean time to silencing one member of an established duplicate pair is approximately $(2d)^{-1} + 10N$ generations, where $N$ is the effective population size. We will approximate the rate of duplicate-gene loss, $L$, by the reciprocal of this time. Then, assuming that a random member of each pre-existing duplicate pair is silenced within each descendent sister species, there is a 50 percent probability that any pair will be divergently resolved. Assuming $b \ll d$, the expected number of duplicate loci in the base population to be divergently resolved after $t$ generations of isolation is

$$\delta_0 = \frac{Gd(1 - e^{-t/2})}{(2d)}, \quad (4.2a)$$

where $G$ is the fundamental number of genes per haploid genome (not including their duplicates).

Subsequent to the isolation of the two lineages, each population will continue to gain $2N_b$ new duplications per locus per generation, each with a probability of fixation equal to 1/$(2N)$ under the assumption that the duplicates are neutral with respect to fitness. Noting again that each new fixed duplication has a 50 percent probability of being resolved in favor of the new locus, the rate of origin of new map changes ultimately becomes $Gb$ per generation. This asymptotic rate of divergence is approached on the time scale necessary for a newly
arisen duplicate (initially in a single copy in a single individual) to first become established by fixation and then to become resolved by nonfunctionalization. To a first approximation, the expected number of map changes resulting from duplication events arising subsequent to geographic isolation is then:

$$\delta_m = Gb \sum_{n=0}^{\infty} (1 - e^{-L})^n = Gb \left( \frac{t - 1 - e^{-L}}{L} \right)$$

and the cumulative number of divergently resolved loci ($\delta$) is given by the sum of eqns (4.2a,b).

Under this model, tens to hundreds of map changes are expected after only a few million generations (Fig. 4.3). Letting $p$ denote the reduction in hybrid fitness per map change, then the expected hybrid fitness following the accumulation of $\delta$ map changes is $W = (1 - p)^{\delta}$. With $p = 1/16$, $W = 0.524$ when $\delta = 10$, and $W = 0.0016$ when $\delta = 100$. With $p = 5/16$, $W = 0.024$ when $\delta = 10$, and $W = 5 \times 10^{-17}$ when $\delta = 100$. Thus, although other factors undoubtedly contribute to species isolating barriers, gene duplication alone appears to be a sufficiently powerful mechanism for the origin of nearly complete genomic incompatibility within a few million years of cessation of gene flow. This is the approximate time scale over which postzygotic isolation generally occurs in animals (Parker et al. 1985; Coyne and Orr 1997).

Genetic theories of speciation have traditionally focused around two competing hypotheses (for reviews, see Orr 1996; Rieseberg 2001). The Dobzhansky–Muller model postulates the accumulation of 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, the general validity of which remains to be demonstrated. For example, the Dobzhansky–Muller model invokes the evolution of co-adaptive complexes of epistatically interacting factors, none of which have been identified at the gene level, whereas the chromosomal model focuses on major chromosomal rearrangements, the fixation of which is greatly inhibited by the reduction in fitness in chromosomal heterozygotes.

A notable feature of the gene-duplication model for speciation is that it is consistent with both the chromosomal and the Dobzhansky–Muller models, while requiring fewer assumptions than either of them. The gene-duplication model is effectively a chromosomal model of speciation, but because the rearrangements are at the microchromosomal level, and hence unlikely to cause significant problems during meiosis, they accumulate passively without any alteration in fitness. The gene-duplication model is also effectively 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 that are simple consequences of reassignments of genes to new locations will appear superficially as epistatic interactions because the loss-of-function phenotype is a function of the number of active alleles at the two homologous loci in hybrid progeny. Thus, genetic analyses of species incompatibilities that fail to identify the specific underlying loci can lead to misinterpretations regarding the underlying genetic mechanism of postzygotic isolation.
Under the gene-duplication model, certain groups of organisms are expected to be more prone to speciation than others, the most notable of which are lineages that experience a doubling in genome size. One potential example of such a key event was noted above—the colonization of ancestral eukaryotic genomes by endosymbiotic organelles, with subsequent variation in transfers of organelle genes to the nucleus being experienced by many of the deeply diverging eukaryotic lineages (Martin et al. 1998). Polyploidization provides another enormous opportunity for the passive origin of isolation via divergent resolution of duplicate genes. Because of the stochastic nature of gene silencing, following the first map changes induced by divergent resolution, thousands of duplicate pairs are expected to remain in a functional state, and these then are free to become divergently resolved in subsequently isolated lineages. As a consequence, ancestral polyploidization has the potential to precipitate a large number of nested speciation events without requiring any underlying changes at the gene level. Adaptive divergence of the isolated lineages may then evolve secondarily. This mechanism for the origin of species may be particularly common in plants, which commonly experience polyploidization events. Indeed, more than a decade ago, Werth and Windham (1991) suggested that divergent resolution following polyploidization may be responsible for the thousands of species of modern-day ferns. Another spectacular example of an adaptive radiation following polyploidization is the relatively young Hawaiian silversword alliance, which is known to have arisen from a combination of genomes of two diploid North American species (Barrier et al. 1999).

Although polyploidization is much less common in animals, key genome doubling events may also have played a significant role in the diversification of the major animal lineages. First, given the apparent three- to fourfold increase in gene content in basal animals relative to fungi, genome amplification might very well have been involved in the origin of the major animal phyla. Second, although the mechanism of genome amplification remains controversial, it also appears that a substantial amount of gene duplication (perhaps equivalent to two polyploidization events) occurred prior to the radiation of the major vertebrate lineages (Postlethwait et al. 2000; McLysaght et al. 2002). Finally, the most speciose lineage of vertebrates, the ray-finned fishes, is a descendant of an ancient polyploidization event, and more recent secondary polyploidization events in fishes appear to be associated with enhanced rates of speciation (reviewed in Taylor et al. 2001).

Summary

Surveys of the contents of completely sequenced genomes provide compelling evidence that gene duplication is an ongoing process in all organisms, with the approximate rate of duplication per gene being of the same order of magnitude as the mutation rate per nucleotide site, if not a little higher. Thus, there is little question that gene duplication plays an important role in genome evolution, and the association of several major phylogenetic radiations with ancient periods of genome amplification suggests an important role in phenotypic evolution as well. The specific mechanisms that are responsible for the preservation and proliferation of duplicate genes are less clear. Although it is tempting to conclude that the cellular complexity of large multicellular organisms required an amplification of gene content relative to that in simpler microbes, these groups of organisms differ in many ways other than phenotypic complexity. Most notable is the tendency for the effective size of a population to decline with an increase in body size. Because random genetic drift is an important determinant of the alternative evolutionary fates of duplicate genes, various aspects of genome-size evolution may be as much an indirect consequence as a cause of organismal complexity. That is, the growth of genome size in large, complex eukaryotic organisms may, in part, be a simple reflection of evolutionary forces that predominate in populations with relatively small effective sizes.

A number of hypotheses have been proposed to explain the long-term preservation of duplicate genes, and an array of comparative analyses point to something other than the evolution of new functions as the predominant mechanism. A substantial
fraction of the data on gene-expression patterns supports the subfunctionalization model, wherein duplicate-gene preservation is driven entirely by the accumulation of degenerative mutations, and most effectively so in small populations. Results from coding-sequence analyses are less conclusive on this issue, but the primary observations are fully compatible with the subfunctionalization model: the fact that the average ratio of replacement to silent substitutions for duplicate pairs is always less than one; the apparent relaxation of selection against amino acid changes in newborn duplicates; and the spatial variation in evolutionary rates exhibited in paralogous copies. Although further functional analysis of duplicate genes and their orthologues in single-copy species will be required to fully evaluate the roles of degenerative and neo-functionalizing mutations in genome evolution, it should be kept in mind that the mutations responsible for the initial preservation of duplicate genes may often be obscured by subsequent mutational refinements to the performance of well-established duplicates. The formal separation of these changes raises considerable challenges for future molecular evolutionary studies.

Despite the remaining uncertainties in our understanding of duplicate-gene evolution, an important central role of degenerative mutations is now firmly established. Although such mutations are normally viewed to be contrary to adaptive evolution, this is not the case with gene duplication. By preserving duplicate genes, subfunctionalizing mutations ensure the continued exposure of both members of a pair to natural selection, thereby increasing the likelihood of occurrence of rare neo-functionalizing mutations. In addition, subfunctionalization provides a potentially powerful mechanism for eliminating the pleiotropic constraints that are unique to single-copy genes, thereby opening up new evolutionary degrees of freedom. Finally, divergent silencing and sub-functionalization of duplicate genes provides a simple passive mechanism for the origin of post-reproductive barriers between isolated populations. Thus, gene duplication may be just as relevant to the origin of new branches in the tree of life as it is to adaptive phenotypic evolution within lineages.