16

The Evolution of Gene Duplicates

Sarah P. Otto*
Department of Zoology
University of British Columbia
Vancouver, British Columbia V6T 1Z4, Canada

Paul Yong
Department of Medicine
University of British Columbia
Vancouver, British Columbia V6T 1Z3, Canada

*To whom correspondence should be addressed: E-mail: otto@zoology.ubc.ca; Telephone: (604) 822-2778; Fax: (604)822-2416.

Advances in Genetics, Vol. 46

0065-2660/02 $35.00

ABSTRACT

Gene and genome duplications have given rise to enormous variability among species in the number of genes within their genomes. Gene copies have in turn played important roles in adaptation, having been implicated in the evolution of the immune response, insecticide resistance, efficient protein synthesis, and vertebrate body plans. In this chapter, we discuss the life history of gene duplications, from their first appearance within a population, through the period during which they rise in frequency or disappear, to their long-term fate. At each phase, we discuss the evolutionary processes that have influenced the dynamics of gene duplications and shaped their ultimate roles within a population. We argue that there is no evidence that organisms have evolved strategies to promote gene duplication in order to permit adaptive evolution. In contrast, many mechanisms exist to silence or eliminate duplicated genes, suggesting that selection has acted largely to reduce the rate of gene duplication. We also argue that natural selection has functioned as an effective sieve, increasing the representation of beneficial gene duplicates among those that establish within a population and that play a long-term role in evolution. To refine our understanding of how selection acts on new gene duplications, we provide a model incorporating a single-copy gene, its gene duplicate, and selection either favoring heterozygotes or eliminating deleterious mutations. Although both forms of selection can increase the initial rate of spread of a gene duplicate, the efficacy with which they do so differs dramatically. Heterozygote advantage always increases the rate of spread and can have a large impact. In contrast, masking deleterious mutations never has a large effect on the rate of spread of the duplicate, and this minor effect can be negative as well as positive. In both cases, the degree of linkage between the two gene copies affects the rate of spread of the duplication. Finally, we discuss evolutionary processes that occur over longer periods after a gene duplication has become established within a population. These long-term processes include maintenance, inactivation, and diversification in function. Consideration of each of the short-term and long-term processes affecting duplicated genes illustrates the subtle ways in which selection has acted to shape genomic structure.

I. INTRODUCTION

The number of genes within the genome will evolve over time whenever (1) there is heritable variation in copy number (caused by gene duplication, gene loss, and polyploidization events) and (2) individuals that carry variant genomes differ in the number of offspring that they bear, either by chance (drift) or as a result of differences in survival ability and/or reproductive potential (selection).

Recent genomic analyses have clarified the extent to which gene duplications (e.g., Lynch and Conery, 2000; Arabidopsis Genome Initiative, 2000) and genome duplications (e.g., Lundin, 1993; Wolfe and Shields, 1997; Postlethwait *et al.*, 1998; Vision *et al.*, 2000) occur. These studies have found that a remarkably high fraction of genes are closely related to other genes within the genome. For example, the fraction of genes that represent recent duplication events (recent enough to generate recognizable paralogs) is 11.2% in *Haemophilus influenzae*, 28.6% in *Saccharomyces cerevisiae*, 65.0% in *Arabidopsis thaliana*, 27.5% in *Drosophila melanogaster*, and 44.8% in *Caenorabditis elegans* (Arabidopsis Genome Initiative, 2000). These numbers increase substantially if less stringent criteria are used to identify paralogs (compare the above results to those of Rubin *et al.*, 2000). Indeed, it seems likely that a large fraction, if not all, genes within a genome are ultimately related by descent to a small number of genes that arose early in our evolutionary history (Maynard Smith, 1998), although there is some evidence for the *de novo* evolution of short regulatory and signal transduction domains (Chervitz *et al.*, 1998). Understanding how these ancestral genes have given rise, through duplication followed by diversification, to the vast number and array of genes present in extant organisms has fascinated evolutionary biologists for decades (Haldane, 1933; Fisher, 1935; Grant, 1963; Spofford, 1969; Ohno, 1970; Stebbins, 1980). In this chapter, we discuss the evolutionary forces that have acted on gene copy number. We briefly review the extent of variation in copy number and the array of mechanisms by which duplications may arise. We then examine processes that affect the initial spread of duplicates within polymorphic populations. Finally, we consider the ultimate fate of these duplicates over longer periods of evolutionary time.

A. Variation in gene copy number

There is great variation in the total number of genes that organisms carry (Table 16.1). Most bacteria have hundreds to thousands of genes, with a minimum near the 480 genes found in *Mycoplasma*. The range is shifted upward by an order of magnitude among the eukaryotes that have been studied. Among metazoans, vertebrates tend to have more genes than invertebrates, perhaps as a result of two genomic doubling events before (~500 MYA) and after (~430 MYA) the divergence of jawless fish (Ohno, 1970; Amores *et al.*,1998; Postlethwait *et al.*,1998; Gibson and Spring, 2000; but see Skrabanek and Wolfe, 1998). Beyond such coarse divisions, there is little relationship between the number of genes carried by an organism and its size or complexity (Valentine, 2000). Indeed, the largest number of genes is likely to be found in organisms that have undergone several rounds of polyploidization. Extensive polyploidization has led to organisms with vast genome sizes, such as the fern *Ophioglossum pycnostichum*, with a record number of 1260 chromosomes in the sporophytic (diploid) phase (Löve *et al.*, 1977).

Table 16.1. Number of Protein-Coding Genes Inferred from Completed Genome Sequencing Projects

Species	Estimated number of genes[a]	Reference
Bacteria		
Mycoplasma genitalum	480	Hutchison *et al.*, 1999
Buchnera sp.	583	Shigenobu *et al.*, 2000
Mycoplasma pneumoniae	677	Tekaia and Dujon, 1999
Rickettsia prowazekii	837	Tekaia and Dujon, 1999
Borrelia burgdorferi	850	Tekaia and Dujon, 1999
Chlamydia trachomatis	894	Tekaia and Dujon, 1999
Treponema pallidum	1,031	Tekaia and Dujon, 1999
Aquifex aeolicus	1,522	Tekaia and Dujon, 1999
Helicobacter pylori	1,577	Tekaia and Dujon, 1999
Haemophilus influenzae	1,709	Fleischmann *et al.*, 1995
Campylobacter jejuni	1,731	Tekaia and Dujon, 1999
Synechocystis sp. PCC6803	3,168	Tekaia and Dujon, 1999
Mycobacterium tuberculosis	3,924	Tekaia and Dujon, 1999
Bacillus subtilis	4,100	Tekaia and Dujon, 1999
Escherichia coli	4,288	Blattner *et al.*, 1997
Archaebacteria		
Methanococcus jannaschii	1,735	Tekaia and Dujon, 1999
Methanobacterium thermoautrophicum	1,871	Tekaia andDujon, 1999
Pyrococcus horikoshii	2,061	Tekaia and Dujon, 1999
Archaeoglobus fulgidus	2,437	Tekaia and Dujon, 1999
Eukaryotes		
Saccharomyces cerevisiae	6,241	Rubin *et al.*, 2000
Arabidopsis thaliana	25,498	Arabidopsis Genome Initiative, 2000
Drosophila melanogaster	13,601	Adams *et al.*, 2000
Caenorhabditis elegans	19,320	C. *elegans* Sequencing Consortium, 1998[b]
Homo sapiens	35,000–120,000[c]	Crollius *et al.*, 2000; Ewing and Green, 2000; Liang *et al.*, 2000

[a]Protein-coding genes.
[b]http://www.sanger.ac.uk/Projects/C_elegans/wormpep/
[c]The number in humans remains so uncertain that geneticists have wagered on its exact value (http://www.ensembl.org/Genesweep/).

It is almost certainly the case that most variation among organisms in gene number simply reflects the predilection for gene or genomic doubling in their ancestors, rather than selection for increased gene number in lineages that are evolving greater complexity. Recently, through genomic analyses of the function and similarity of genes, it has become possible to estimate the number of genes in the "core" proteome. The proteome is defined as the number of protein-coding

genes that are so divergent from one another that they no longer share sequence similarity, which uses the fairly arbitrary benchmark of whether or not we can detect sequence similarity to assess uniqueness. One might expect that counting genes in this way would lead to a strong relationship between gene number and complexity, but this also appears to be false. For example, *Drosophila* has a core proteome of 10,736 (8065) genes, while C. *elegans* has a core proteome of 14,177 (9453) genes (Arabidopsis Genome Initiative, 2000; second estimates in parentheses from Rubin *et al.*, 2000), despite the fact that flies have roughly six times more cell morphotypes than nematodes (Valentine, 2000). Again, the size of the core proteome likely reflects the past tendency toward gene or genome duplication among ancestors distant enough that the duplicated genes have diverged substantially in sequence.

B. Mechanisms of increase in gene copy number

Gene copy number has increased through both duplication of relatively small regions of DNA and through genome-wide duplication events. The mechanisms involved in these processes are quite different, and we briefly review them in turn.

1. Gene duplication

Stretches of DNA may be duplicated in a variety of ways, although the main mechanisms involve mobile elements and unequal crossing over. Replicative transposons, retrotransposons, and retroviruses are elements that encode proteins enabling their own replication and insertion into the host genome. Occasionally, these elements replicate neighboring host genes as well as their own (Lewin, 1994; Baldo and McClure, 1999). Although such "accidents" may be rare per replication event, transposition can occur frequently enough to ensure a substantial rate of production of gene duplicates. Indeed, Charlesworth and Langley (1991) note that, in *Drosophila*, replicative transposition occurs at a rate of approximately 10^{-4} per transposable element per generation, which is orders of magnitude higher than the point mutation rate of a transposable element.

Duplications can also arise through unequal random breakage and reunion between nonhomologous sequences (Maeda and Smithies, 1986). An example is the human haptoglobin genes, where Hp2 is a result of what appears to be an unequal exchange between the fourth intron of Hp1F and the second intron of Hp1S. However, the crossover points show no homology, providing evidence that an unequal nonhomologous breakage and reunion has occurred. More frequently, unequal homologous recombination may occur between repeated elements, including transposable elements (Charlesworth and Langley, 1991), causing either the duplication or loss of intervening sequences. A signature of such

a recombination event is a copy of the repeated element between the gene and its duplicate. An example is the human growth hormone and human chorionic somatomammotropin (HGH–HCS) gene cluster, where a SINE sequence resides in between the two genes. The rate of unequal crossing-over can rise dramatically in areas of tandemly repeated genes. For example, Guillemaud *et al.* (1999) examined the rate of unequal recombinants in a tandemly repeated array of *esterase* B genes in *Culex pipiens*, finding new gene amplifications in 4 of 60 mosquitoes (7%). This represents an extremely high rate of gene duplication compared to studies of regions lacking tandem arrays. In *D. melanogaster*, for example, only nine tandem duplications were detected in the *rosy* region (Gelbart and Chovnick, 1979), among 2.25 million offspring examined. We conclude that the mechanisms by which duplicate genes are generated vary greatly, depending on the activity of mobile elements and on the nature of the surrounding sequence.

2. Genome duplication

In addition to gene duplication, whole-scale genome duplication (polyploidization) has been a major source of gene duplicates. The main mechanisms by which polyploidization is achieved have recently been reviewed by Ramsey and Schemske (1998). Typically, polyploidization is the result of an error that occurs at one of three points: (1) mitosis, (2) meiosis, or (3) fertilization (Stebbins, 1980; Ramsey and Schemske, 1998; Otto and Whitton, 2000). Failure of a mitotic cell division can result in polyploidization and is known to occur in both animals and plants. If the failure occurs within the germline or, more generally, within cells that give rise to reproductive tissue, then the polyploid genome may be inherited. However, failure of the first or second reductive division during meiosis is a more usual route to polyploidy. Ramsey and Schemske (1998), reviewing several plant studies, estimated that unreduced gametes comprise 0.56% of pollen from nonhybrid individuals and as much as 27.52% of pollen in hybrids. Errors in fertilization are also a common source of polyploids. Fertilization of an egg by multiple sperm (polyspermy) is the most widespread error in fertilization; in humans, for example, polyspermy is the main mechanism by which triploids (~1–3% of conceptions; McFadden *et al.*, 1993) are produced (Uchida and Freeman, 1985). Another error in fertilization that can occur among asexual species is the occasional fertilization of an unreduced egg by sperm from a related sexual species. This is a likely explanation for the unusually high incidence of polyploidy among animals that are gynogenetic (in which fertilization requires stimulation by sperm that are not incorporated genetically) or hybridogenetic (in which sperm contribute genetically to the zygote, but the paternal genome is lost sometime during development; Vrijenhoek *et al.*, 1989; Otto and Whitton, 2000).

II. EVOLUTIONARY FORCES AFFECTING GENE COPY NUMBER

Given that gene and genome duplications do arise, we now turn to an examination of how evolutionary forces, including selection, mutation, and genetic drift, shape the evolution of gene number. We first discuss the rate at which duplications occur; in particular, we ask how this rate may have been molded by evolution. Second, we examine the processes that affect the frequency of a gene duplicate after it arises. Selection during this period has a very strong influence on whether the duplicate is lost or rises in frequency within the population. Finally, we turn to long-term evolutionary processes that act on a gene duplicate after it is established in a population. These long-term processes determine the ultimate fate of the gene duplicate.

A. The rate of duplication

The completion of several genome sequencing projects has enabled researchers to quantify, in an accurate and unbiased manner, the extent to which the genome is comprised of duplicated sequences. The enormity of the contribution of duplications has recently been assessed in *Arabidopsis thaliana:* 17% of its genome falls within one of 1528 tandemly duplicated arrays, while an even greater percentage (~58%) falls within one of the 24 large duplicated segments that most likely arose via polyploidization (Arabidopsis Genome Initiative, 2000). These numbers are limited to detectable and thus relatively recent events (within the past 200 MY or so; Vision *et al.*, 2000) and hence underestimate the total fraction of genes that have arisen through duplication.

1. Rate of gene duplication

What is the rate at which new gene duplications appear? Lynch and Conery (2000) recently addressed this question by scouring complete genomic sequences for very young gene duplicates (excluding multigene families and transposable elements). They found 10 pairs of genes in *D. melanogaster*, 32 pairs in *S. cerevisiae*, and 164 pairs in *C. elegans* whose level of silent-site divergence was less than 0.01. Using estimated rates of silent-site substitution, they inferred that 31 new duplicates arose per genome per million years in flies, 52 in yeast, and 383 in nematodes. In the absence of selection, one can use the neutral theory (Kimura, 1983) to make two important predictions about gene duplications from these estimates. (1) The rate at which gene duplicates spread throughout a population (= a substitution) should equal the genome-wide rate of gene duplication, ν, regardless of the population size. (2) The probability that any two haploid genomes randomly sampled

from a diploid species will differ in terms of the duplicate genes that they carry is $4N_e\nu/(4N_e\nu + 1)$. (Aside: N_e is a measure of the size of a population in terms of how much allele frequencies change by chance from generation to generation, i.e., as a result of random genetic drift. A population with a very large census size, N, may nevertheless have a low N_e if only a few individuals ever reproduce. See Crow and Kimura, 1970, for further details.) Given their estimates of ν, Lynch and Conery (2000) estimate that roughly half of the genes within the genome are expected to produce duplicates that spread through the population over a time period of 35–350 million years. For scale, mammals and birds first appeared in the fossil record around 200–250 MYA, which is also the time period during which monocots and dicots diverged. From the second prediction of the neutral theory, we expect that a large fraction of individuals in sizeable populations ($N_e > 10^4$) should be heterozygous for a duplication somewhere within their genome. It is, however, unreasonable to assume that new gene duplicates are always neutral (see Section II.B). If duplicated genes are often immediately deleterious, lineages that have survived to be sequenced would have fewer accumulated duplications than expected from the neutral theory. As an extreme example, if all gene duplicates were lethal in the heterozygous condition, then surviving individuals would never carry a gene duplicate, regardless of the rate at which gene duplicates arise. Consequently, the method of Lynch and Conery would underestimate the rate of gene duplication to the extent that selection has eliminated deleterious duplicates.

2. Rate of genome duplication

In addition to high rates of gene duplication, genome duplication has occurred repeatedly, especially in plants. A large fraction of angiosperms are thought to have undergone genomic doubling at some time in their evolutionary history, with estimates ranging from 20–40% (Stebbins, 1938), to 57% (Grant, 1963), to 70% (Goldblatt, 1980; Masterson, 1994), depending on the method used to infer polyploidization events (reviewed in Otto and Whitton, 2000). These estimates are insufficient to gauge the role of polyploidization in genome evolution, however, because they tell us only whether polyploidy has occurred at some point in the past. Groups that have polyploidized once and those that have done so repeatedly are treated equally. Recently, we have developed a new method to estimate the rate of polyploidization based on the excess of species bearing an even number of chromosomes at the gametic or haploid stage. Using this method, we inferred a rate of polyploidization per speciation event of ~2–4% in angiosperms and ~7% in ferns (Otto and Whitton, 2000). Even though polyploidy is much rarer among animals, we identified over 170 independent polyploidization events among insects and vertebrates, with many more known from other invertebrate species (see citations in Otto and Whitton, 2000). Several of these events

represent ancient polyploidization events and affect the number of gene copies in a large fraction of animals. Large taxonomic groups of polyploid fish include the Actynopterygii (the highly speciose ray-finned fishes), Catostomidae (the sucker family), Salmonidae, the group *Corydoras–Aspidoras–Brochis* (Callichthyidae, the catfish family), the subfamily Schizothoracinae (Cyprinidae, the carp family), the subgenus *Barbus* (Cyprinidae), and the subgenus *Labeobarbus* (Cyprinidae). Among amphibia, polyploid groups include the Sirenidae and *Xenopus* (Pipidae). Finally, the extensive collinearity of genetic content on different chromosomes within vertebrates (Postlethwait *et al.*, 1998; Gibson and Spring, 2000) supports Ohno's (1970) contention that vertebrates underwent two genomic doubling events early in the paleozoic era.

3. Evolution of the rate of duplication

Even though duplication has played a large role in evolution, it is a logical error to assume that organisms have necessarily evolved strategies to promote duplication per se. Another possibility is that the rate of duplications has evolved to its current level as a side consequence of selection on other aspects of the replicative machinery of the cell, including selection to reduce the rate of point mutations, selection to ensure proper chromosome segregation, and selection to reduce the energy and time costs of DNA replication. Simply put, duplication events may be the consequence of the imperfection of cellular processes involved in DNA replication, and the rate of these errors might reflect a balance of selective forces other than the fitness effects of gene duplications. To demonstrate that organisms have evolved strategies to promote or hinder gene duplications, one must be able to reject the null hypotheses that they have not. One way to reject this null hypothesis is to show that mechanisms have evolved specifically to increase or decrease the rate of gene duplication.

In fact, complex mechanisms have evolved by which genes can promote their duplication, including replicative transposition, reverse transcription, and insertion. These events are made possible by a suite of enzymes, including transposase, resolvase, reverse transcriptase, and integrase (Lewin, 1994), which are encoded by the very elements that become duplicated but otherwise play no essential role in the cell. Nevertheless, these mechanisms almost certainly evolved to promote the duplication of certain genetic elements (transposons, retroposons, and retroviruses) and not as part of a generalized strategy to promote gene duplication on the part of an organism. This claim is supported by the fact that these mechanisms do not duplicate all possible gene sequences. For example, integrase requires a conserved CA near the end of an inverted repeat to allow insertion into the host genome of DNA transcribed from the RNA of retroposons and retroviruses (Lewin, 1994). A generalized mechanism to promote gene duplication would not be so picky.

Conversely, complex mechanisms have evolved to decrease the rate at which gene duplications become established within populations (reviewed by Selker, 1997, 1999, 2002; Wu and Morris, 1999). The most remarkable of these mechanisms is repeat induced point mutation (RIP), whereby duplicated genes are recognized by a poorly understood process and are subsequently subject to a high rate of mutation, which generally inactivates both gene copies. RIPing is only known to occur in *Neurospora,* but several other mechanisms that either recognize and silence duplicated sequences or degrade messenger RNA transcripts of duplicated genes have been uncovered, including methylation induced premeiotically, quelling, transvection, and paramutation (see Wu and Morris, 1999, for a description of these and other homology-dependent silencing processes). In many cases, these mechanisms were discovered because they interfered with expected patterns of expression in transgenic organisms. For example, the *hygromycin phosphotransferase* gene (*hpt*) was introduced into *Arabidopsis*. Initially, this transgene conferred resistance to hygromycin, but this resistance was lost in multiple homozygous lines as a result of duplicate gene silencing without loss of the transgene (Mittelsten Scheid *et al.*, 1991). Many of these mechanisms appear to be taxonomically widespread and general in action, although duplicated genes do vary in their susceptibility to silencing, depending on their sequence and site of insertion (e.g., inverted repeats appear to be particularly prone to silencing; Selker, 1999). It is thought that these homology-associated mechanisms have evolved as defense mechanisms against the proliferation of transposable elements and retroviruses (e.g., Ratcliff *et al.*, 1997; Al-Kaff *et al.*, 1998). The facts that many eukaryotic genomes are riddled with transposable elements (~35% of the human genome may be relics of mobile elements; Wolffe and Matzke, 1999) and that transposition may induce severe mutations (insertions and chromosomal rearrangements) suggest that there may well have been a substantial selective advantage to inactivating such elements.

Similarly, there is no evidence that organisms have evolved mechanisms to promote polyploidization in order to increase the number of genes available for evolution to act upon. As discussed earlier, polyploidization is typically the result of an error in mitosis, meiosis, or fertilization. There are some very unusual reproductive systems in which polyploidization occurs frequently and regularly, suggesting that some mechanism has evolved to promote polyploidization. For example, in some soft-scale insect species (Coccidae: Homoptera: Insecta), males develop from unfertilized haploid eggs that then polyploidize by the fusion of the first two cleavage nuclei (Nur, 1980). Even in these cases, however, the mechanisms promoting polyploidization have not been selected to increase the number of genes, for in fact this number stays constant from one generation to the next. Instead, polyploidization has evolved simply to restore the diploid state without fertilization. On the other hand, it is commonly observed that genome size declines after polyploidization events and that gene expression patterns of polyploid

lineages decline toward diploid levels, a process known as rediploidization (Ferris and Whitt, 1977; Grant, 1981; Leipoldt, 1983; Werth and Windham, 1991; Soltis and Soltis, 1993). These observations suggest that mechanisms may have evolved to eliminate genes added by polyploidization. Although rapid and extensive genomic rearrangements and gene silencing or loss can occur following polyploidization (reviewed by Soltis and Soltis, 1993, 1999), this process is poorly understood. Mittelsten Scheid *et al.* (1996) observed gene silencing of a single-copy transgene after a change in ploidy, which they argue is not a homology-dependent process but one that depends on chromosome number or genome size. More work on the mechanisms by which gene silencing and loss occurs following polyploidization promises to clarify several puzzling observations, such as rediploidization.

Thus, the preponderance of evidence suggests that organisms have evolved processes to hinder rather than to promote the establishment of gene duplicates within their genomes. Consequently, we can reject our null hypothesis, but not in favor of the hypothesis that organisms have evolved strategies to promote the generation of gene copies. Rather, we suggest that organisms have evolved strategies to silence or inactivate gene duplicates before they rise to fixation within a population. Why then are duplications so widespread? We argue that the costs of ensuring error-free meiosis, mitosis, fertilization, and recombination would prevent the rate of gene or genome duplication from ever evolving to zero. Similarly, mechanisms that recognize and silence gene duplications must entail several costs, including the time involved in searching out homologous sequences, the energetic costs of silencing (including the costs of producing enzymes that promote silencing), and the risk of silencing critical genes and gene families. If the benefits of eliminating or silencing duplicated genes are not always greater than these costs, then homology-dependent silencing mechanisms may evolve to a point where the benefits and costs balance. Further reductions in the rate of establishment of gene duplicates would then be disadvantageous. One prediction from this line of reasoning is that we would expect stronger homology-dependent gene silencing mechanisms in organisms with a history of infectious spread of mobile elements, because the past benefits to silencing would have been greater, shifting the balance point to a higher overall level of homology-dependent silencing. A similar trade-off is thought to play an important role in the evolution of mutation rates (Sniegowski *et al.*, 2000).

In summary, gene and genome duplication events are commonplace. There is currently no good evidence that organisms have evolved mechanisms to increase the chance that their offspring carry gene or genome duplicates, although mobile elements within organisms clearly have evolved mechanisms to promote their own replication. In contrast, evidence from several species suggests that a variety of mechanisms have evolved to eliminate or silence gene duplicates soon after they arise. Thus, there is every reason to believe that the appearance of duplicated segments within genomes represents nothing more than a series of historical

accidents. What happens to these duplicated genes after their appearance is the subject of the next section.

B. The spread of new duplicates

After a new gene duplicate has appeared within a population, its evolution may be divided into two distinct phases, a polymorphic period and a fixed period (Ohta, 1988c). Although evolutionary forces such as selection and mutation may affect both phases, their consequences are different. We consider first the polymorphic period and will return to the fixed period in the following section. During the polymorphic period, the gene duplicate, which is originally present in a single copy, changes in frequency, ultimately becoming lost or fixed within the species. In humans, such polymorphic gene duplicates have been found in both the opsin (Wolf *et al.*, 1999) and ribosomal (Veiko *et al.*, 1996) gene families. If the gene duplicate were selectively neutral, it would have a $1/(2N)$ chance of becoming fixed within a diploid population of census size N, and it would take on average $4N_e$ generations to do so. These estimates assume that gene silencing, mutations, and unequal crossing-over events that silence, inactivate, or knockout the duplicate are rare ($\ll 1/2N$ in frequency); if such events are frequent ($\geq 1/4N$), it is essentially impossible for the entire population ever to contain two active gene copies (proven using equation 8.8.38 of Crow and Kimura, 1970). This assumes that the gene duplication arises only once. In a series of papers, Ohta (1987, 1988b, 1988c) developed models in which duplicates are continuously being produced and destroyed by unequal crossing over. She also found that gene families are unlikely to evolve when duplicate genes are inactivated at an appreciable rate, unless there is positive natural selection favoring duplicates carrying different alleles (Ohta, 1987).

As noted by Walsh (1995), it is often assumed that gene duplications rise to fixation by random genetic drift, such that the gene duplications observed in a genome are a random subset of all the duplications that have occurred. However, selection acting on individuals carrying a duplicated gene has a profound effect on the probability of fixation of the duplicate. If the duplicate alters the fitness of its heterozygous carriers by a factor $1 + s$, through changes in survival, mating ability, or reproductive potential, its fixation probability becomes:

$$\frac{1 - e^{-2sN_e/N}}{1 - e^{-4sN_e}} \tag{1}$$

(Crow and Kimura, 1970). To simplify this discussion, we assume that the population is diploid and that the duplicate is additive in action, so that a homozygote carrying the duplication has an expected fitness of $1 + 2s$ (see Crow and Kimura, 1970, equation 8.8.3.21, to incorporate dominance). As equation (1) confirms, the probability of fixation is always lower than $1/(2N)$ if the gene duplicate

reduces fitness (i.e., $s < 0$). Conversely, if the duplicate has a beneficial effect on fitness, its probability of fixation is greater than $1/(2N)$, being approximately $2s$ for small s. Notice that even directly beneficial gene duplicates are not guaranteed to fix within a population; 98% of gene duplicates that confer a 1% fitness advantage will ultimately be lost by chance!

It seems likely that when a gene duplicate does affect fitness, it will most often have a negative effect ($s < 0$). Gene duplicates may insert into other genes or gene regulatory regions, disrupting their function. Gene duplicates may also induce deleterious chromosomal rearrangements as a result of ectopic exchange. Gene duplicates may have been incompletely or inaccurately copied in such a way that the gene product decreases fitness. For instance, gene duplicates that have been reverse transcribed from a processed RNA intermediate may lack introns that regulate gene expression or coordinate alternative splicing, leading to an incorrect expression pattern of the gene products. Furthermore, the timing or level of gene expression may be disturbed by the presence of different regulatory factors near the new position of the duplicated gene. Even when both gene copies are properly expressed, there may be a fitness cost to having increased levels of the gene product. The opposite problem, underexpression, may also arise if the gene duplicate induces inactivation or silencing of both the original and the duplicate copy, as is observed with RIPing and quelling (Selker, 1997). Of course, each such change may increase rather than decrease fitness, but the latter will be more common as long as the duplication causes a relatively large change in gene expression and the original unduplicated gene was functioning well in its current environment. Under these circumstances, the chance that an alteration in gene expression will improve the ability of an individual to survive and reproduce is small. Such an argument applies to mutations of any form and has received both theoretical validation (Fisher, 1930; Kimura, 1983) and empirical support (Mukai *et al.*, 1972; Simmons and Crow, 1977; Deng and Lynch, 1997; Elena, 1997; Vassilieva *et al.*, 2000, but see Shaw *et al.*, 2000).

Even if a small fraction of gene duplicates that arise are beneficial, the subset of duplicates that fix within a population will contain a much higher fraction that increase fitness (Figure 16.1). Typically, deleterious gene duplicates are lost soon after they arise, neutral duplicates only rarely fix, and beneficial gene duplicates have a much larger chance of becoming established within a population. Consequently, gene duplicates that remain within a genome for long periods of time are a very biased subset of the ones that have occurred, a phenomenon known as the selective sieve. This helps explains, for example, why the number of copies of each transfer RNA in *C. elegans* correlates with codon usage in highly expressed genes (Duret, 2000). The simplest explanation for this is not that the tRNAs most in demand were duplicated at a higher rate, but that they were more likely to rise to fixation when they did occur.

Although it may be rare that a new gene duplicate confers a fitness advantage, beneficial duplicates do occasionally arise. The most likely source of a

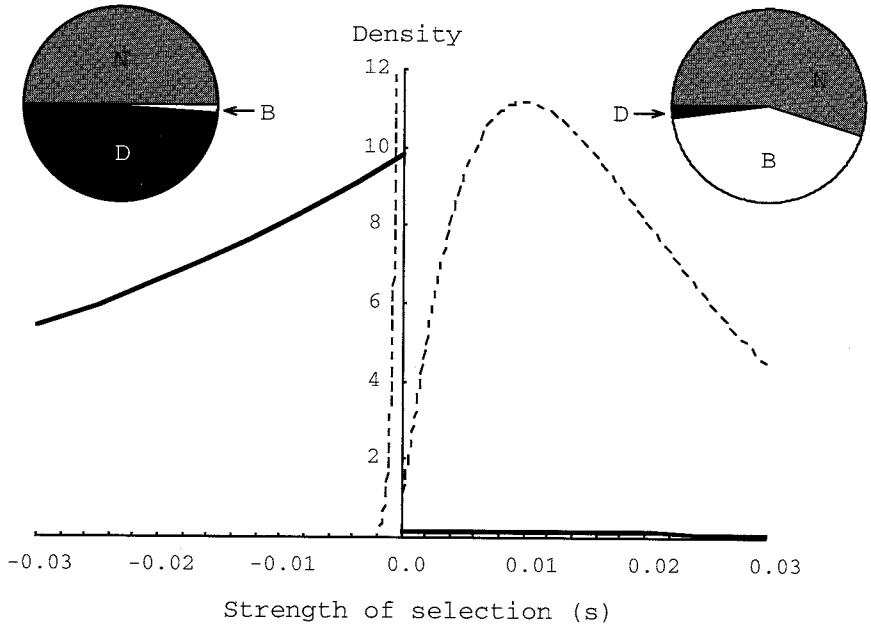

Figure 16.1. The selective sieve. Because selection affects the probability that a mutation will establish within a population, the effects of new gene duplicates on fitness will be very different from the effects of those gene duplicates that fix within a population. In the central figure, a probability density function for the selection coefficient (s) is illustrated, with the thick curves showing a hypothetical probability distribution among new duplicates, and the dashed curves showing the distribution among fixed duplicates (i.e., weighted by equation 1). It is assumed that the strength of selection follows a negative exponential distribution for both deleterious and beneficial duplicates, with the average magnitude of selection equal to 0.01. Fifty percent of new duplicates are assumed to be neutral (N), 49% deleterious (D), and only 1% beneficial (B), as illustrated in the left pie chart. Among fixed duplicates, however, these proportions change radically (right pie chart), with a full 43% of duplicates having a selective advantage. A population size of 1000 is assumed. The selective sieve would be even more effective in larger populations or with a larger average strength of selection.

fitness advantage is that, prior to duplication, there was insufficient gene product relative to the demands of the organism in its current environment. Of course, point mutations that increase expression would also have been advantageous under these circumstances, but gene duplications may be a relatively frequent mutation affecting expression levels. Such a scenario is thought to explain the extensive duplication of both transfer and ribosomal RNAs (Ohno, 1970). It is also thought that selection to increase enzyme production favored the spread of duplicates of *glutathione S-transferase* genes in houseflies (Wang *et al.*, 1991) and *esterase* genes in the mosquito, *Culex* (Mouchès *et al.*, 1986; Guillemaud

et al., 1999), both of which confer resistance to insecticides. Similarly, selection for metal tolerance may be related to a high incidence of *metallothionein* gene duplicates (Maroni *et al.*, 1987). If the benefit of gene duplicates often stems from selection for increased gene product, one would predict that the duplication of highly expressed genes is more likely to be advantageous, assuming that expression levels are correlated with demand for gene product. This prediction was confirmed in yeast, where genes with high mRNA levels, including heat shock, glucose metabolic, and cytosolic ribosomal genes, were more likely to be found in duplicate than genes with low levels of expression (Seoighe and Wolfe, 1999).

Another advantage to gene duplication may occur when carrying two different alleles of a gene is beneficial, i.e., when there is heterozygote advantage (Spofford, 1969; Ohno, 1970; see also Appendix). In this case, a gene duplicate carrying one allele can spread along with an alternative allele at the original gene, generating permanent heterozygosity. A striking example has been studied in *Culex pipiens* and again involves selection for insecticide resistance (Lenormand *et al.*, 1998). Organophosphate insecticides (OPs) target acetylcholinesterase (AChE), an enzyme that hydrolyzes acetylcholine and plays an important role in the normal functioning of the central nervous system. Mutations modifying the target of OPs led to resistant alleles ($Ace.1^R$), which have a higher fitness in the presence of OPs but which lower fitness in their absence due to the reduced activity of the $Ace.1^R$ allele. A gene duplication appeared that combined both resistant and sensitive alleles ($Ace.1^{RS}$). $Ace.1^{RS}$ individuals were found to be almost as resistant to insecticides, but they suffered from a much lower fitness cost because of the normal functioning of the sensitive allele. The duplication spread rapidly in the south of France, with an estimated selective advantage over $Ace.1^R$ of 3–6% (Lenormand *et al.*, 1998). In this case, the gene copies are tightly linked and the haplotype contains both alleles, a combination that we show in the Appendix is especially conducive to the spread of the gene duplicate when there is heterozygote advantage.

An often touted, but much less straightforward advantage, of gene duplication is that it can provide backup copies that protect an individual from either heritable (mutations) or nonheritable (developmental) errors. A critical point to remember, however, is that the benefit of such masking is roughly equal to the chance that an error occurs times the fitness consequence of the mutation or developmental error (Pál and Hurst, 2000). Thus a mutant allele or a developmental error that occurs in only one in a million individuals can account for only a miniscule benefit to gene duplication (s is proportional to 10^{-6}). Another problem pointed out by Pál and Hurst (2000) is that the majority of nonheritable errors should have minor fitness consequences, because they would affect only a fraction of the expected gene products within the organism (e.g., a translational error would affect only one protein within one cell). They argue that the most

damaging errors are those involving gene silencing, which can be propagated to daughter cells, but they note that such errors are less common. A more substantial problem exists with the hypothesis that gene duplicates protect an individual from deleterious mutations: they also double the mutational target and reduce the efficiency by which selection can eliminate deleterious mutations by masking their effects. The result is that the frequency of mutant alleles rises in the population, especially among those individuals carrying the gene duplication. Clark (1994) showed that if deleterious mutations are fully redundant and if the two gene copies are completely linked, then the advantage of masking and the disadvantage of accumulating mutations exactly balance, so that the duplicate is no more likely to spread when rare than if it were neutral. In the Appendix, we generalize this model to allow for partial redundancy and for arbitrary linkage between the original and duplicate gene. We find that, unless there is perfect redundancy (i.e., unless mutations are completely recessive), tandem duplications are actually selected against because deleterious alleles remain associated with the gene duplicate for longer periods of time. In contrast, loosely linked duplications can gain a positive selective advantage from masking as long as the fitness of individuals carrying a deleterious mutation is sufficiently higher when a functioning duplicate gene is present than when it is not (see equation A7). Nevertheless, the magnitude of this selective advantage is never large (always being less than twice the mutation rate). Thus, masking deleterious mutations will have only a minor effect on the probability of fixation of a gene duplication (see equation 1).

C. Maintenance and divergence of gene duplicates

Once a gene duplicate has fixed within a population, its evolution is by no means over. In general, three eventual fates exist for duplicated genes: (1) maintenance of both copies performing the exact same functions (Figure 16.2B), (2) inactivation or loss of one copy (Figure 16.2C), or (3) divergence in function through neo-functionalization (Figure 16.2D), specialization (Figure 16.2E), or subfunctionalization (Figure 16.2F). Most empirical data on the likelihood of each of these fates comes from studies of ancient polyploids. A surprisingly high proportion of genes that duplicated via polyploidization have been retained over long periods of time: ~8% in yeast over ~100 MY (Seoighe and Wolfe, 1999), ~72% in maize over ~11 MY (Ahn and Tanksley, 1993; Gaut and Doebley, 1997), ~77% in *Xenopus* over ~30 MY (Hughes and Hughes, 1993), ~70% in salmonids over 25–100 MY (Bailey *et al.*, 1978), ~47% in catastomids over ~50 MY (Ferris and Whitt, 1979), and ~33% in vertebrates over ~ 500 MY (Nadeau and Sankoff, 1997). Many of these duplicates have diversified in function and, more typically, in expression patterns (e.g., Ferris and Whitt, 1979). Additional data on the fate of duplicated genes was also provided by the genome-wide analyses by Lynch and Connery (2000). They noted that there is a nearly exponential decline in the number of functional duplicates over time, where the time of each duplication

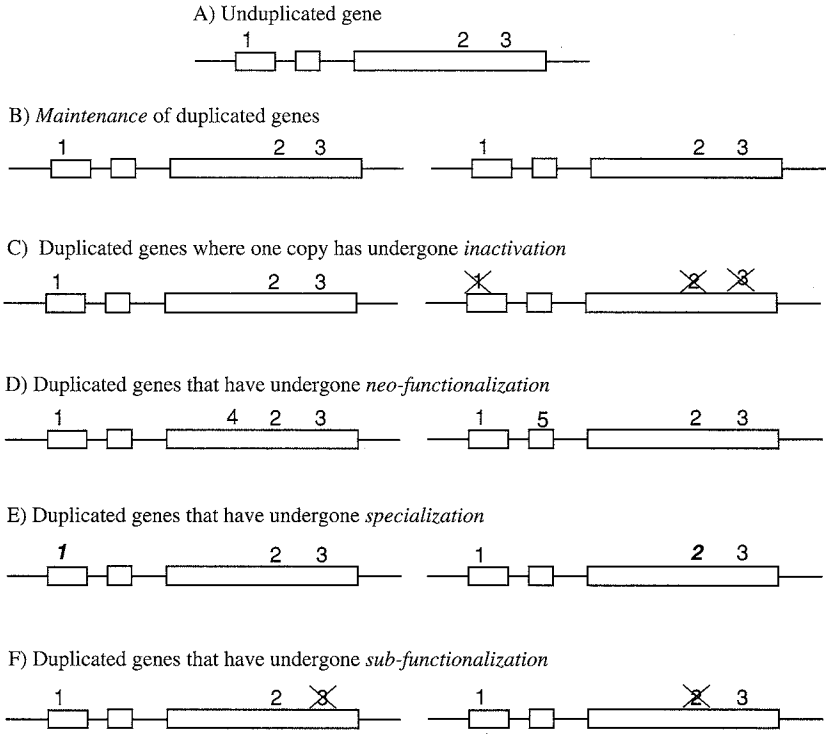

Figure 16.2. The ultimate fate of a gene duplication (for further details, see Force *et al.*, 1999). (A) A single-copy gene with three exons (boxes) is shown. For illustration purposes, we imagine that there are sites (1, 2, 3) that are critical for the performance of three different functions. These may be sites where regulatory proteins bind or where there is an active site in an enzyme. We consider the ultimate fate of a gene duplication that preserves the structure and function of the gene. (B) Both copies may be maintained indefinitely. (C) One copy may be inactivated by the accumulation of mutations that disrupt its functions (X). (D) The two copies may diverge and take on novel functions ("neo-functionalization"; 4, 5). (E) The two copies may specialize in function, each becoming better at performing one of the original functions of the single-copy gene **(1, 2)**. (F) The two copies may accumulate deleterious mutations disrupting different functions ("subfunctionalization"), causing both to be maintained in order to perform all the functions of the original single-copy gene.

event is estimated by the number of silent site differences between the duplicate copies. Assuming a molecular clock, they infer an average half-life of duplicate genes of about 4 million years, which is a much higher inactivation rate than observed following polyploidization.

A large body of theory has been developed to examine which of the three ultimate fates is most likely and under what conditions (Bailey *et al.*, 1978; Takahata and Maruyama, 1979; Watterson, 1983; Walsh, 1987; Ohta, 1988a;

Hughes, 1994; Walsh, 1995; Nowak *et al.*, 1997; Force *et al.*, 1999; Lynch and Force, 2000). We turn now to a summary of these ideas.

1. Maintenance of function

Whether two gene copies can be maintained over long periods of evolutionary time depends strongly on whether the genes are redundant in function. If the two gene copies are perfectly redundant, i.e., if a single functional copy can completely mask the deleterious effects of mutations in other copies, they are very unlikely to be maintained (Fisher, 1935; Cooke *et al.*, 1997). With complete redundancy and total overlap in function, Fisher (1935) showed that the maintenance of both copies occurs only if one gene performs its function less efficiently and is less mutable than the other gene. Fisher also noted that both copies could be maintained if they had the exact same efficiency and number of mutations, but this would only work in an infinitely large population. Continual gene conversion, if frequent enough, could also ensure that both copies are maintained indefinitely (Walsh, 1987). On the other hand, the long-term maintenance of both gene copies is much more likely whenever redundancy is incomplete, such that mutations are deleterious even when they occur in only one gene copy (Takahata and Maruyama, 1979). In this case, selection acts against changes at critical sites in either copy, although the copies would diverge in sequence at silent sites. Such purifying selection has been observed in 17 duplicates pairs of genes in the anciently tetraploid *Xenopus laevis*, where Hughes and Hughes (1993) found that amino acid change was rare relative to the rate of silent substitution in both copies, suggesting that neither copy of the gene has been free to accumulate amino-acid altering mutations. Given that favorable duplicates comprise a much higher fraction of duplicates that fix within a population than their representation among new duplicates (as a result of the selective sieve; Figure 16.1), it is more likely that purifying selection will act to preserve the duplicates that have fixed. Thus, for example, the maintenance of 579 tRNAs within the *C. elegans* genome (Duret, 2000) may be explained by the need for a large pool of tRNAs within the cell, such that loss of a tRNA is deleterious even though functioning duplicates of the same tRNA exist within the genome. Another issue to consider is the source of the duplicated genes. Following a polyploidization event, the relative ratios of most (but not all; Galitski *et al.*, 1999) gene products remain roughly the same. Hence, the loss of a gene that has been duplicated by polyploidization may actually cause an underproduction of its gene product relative to those with which it interacts. Thus, selection to maintain gene copies may be stronger when the copies are produced by polyploidy than by gene duplication, which may help explain the apparent paradox that gene duplicates that arise following polyploidization appear to be retained for longer than gene duplicates that arise individually (Lynch and Conery, 2000).

2. Inactivation

If the integrity of the two gene copies is not actively maintained, the most likely fate of a gene duplicate is extinction, i.e., inactivation or loss (Haldane, 1933; Bailey *et al.*, 1978; Takahata and Maruyama, 1979; Watterson, 1983; Walsh, 1995). When there is complete redundancy, Takahata and Maruyama (1979) showed that extinction is extremely rapid; typically, the half-life of a duplicate was less than 10 times the population size in generations. They concluded that their results are inconsistent with the percentage of gene duplicates that have been retained following polyploidization in salmonids and catostomids. Their model, however, only allowed deleterious mutations to accumulate in each gene. Walsh (1995) included both mutations that inactivate a gene and mutations that cause a gene duplicate to take on a novel function. Nevertheless, Walsh found that inactivation remains a very likely outcome, unless mutations causing novel functions are extremely common and the population size is large. For example, if mutations are 20,000 times more likely to inactivate a duplicate than they are to provide a novel function, then even if this novel function confers a 1% fitness advantage, inactivation occurs before neo-functionalization 99% of the time in a population with an effective size of 5000. Inactivated gene duplicates (pseudogenes and/or poorly expressed gene copies) have been observed in many gene families (e.g., the MHC class I*b* genes; Hughes and Nei, 1989). Such examples of inactivation are characterized by a high rate of amino acid substitution, suggesting that purifying selection is weak or absent.

3. Divergence in function

Nevertheless, many gene duplicates do diversify in function over time, taking on novel functions, specializing in functions originally shared by the single-copy gene, or differentiating in the timing and/or location of expression (Figure 16.2D–F). Hughes (1994) reviewed several examples in which evidence has been found for positive selection having acted to diversify gene duplicates. Hughes also suggested that the chance of diversification might be substantially higher whenever the ancestral gene performed multiple functions, which he termed "gene sharing." In this case, selection might actively favor the specialization of each gene copy on different subsets of these functions. As an example, he cites δ-crystallin, a gene that encodes both an enzyme (argininosuccinate lyase) and an eye-lens crystallin in ducks; in other vertebrates, however, this gene appears to have duplicated such that each gene encodes only one of these products. Recently, it has been noted that positive selection for specialization is not necessary for this diversification process to occur. Force and colleagues (Force *et al.*, 1999; Lynch and Force, 2000) noted that inactivation may itself lead to specialization. That is, if the original gene had several functional or regulatory domains, deleterious mutations disrupting

different subfunctions could accumulate in each of the gene copies (Figure 16.2F). Subsequently, both gene copies would be essential and could not be further inactivated (see also Werth and Windham, 1991). This process, which they call "subfunctionalization," preserves the duplicates from loss or full inactivation over longer periods of evolutionary time, which increases the opportunity for adaptive mutations to occur. The fact that duplicate genes often differ in the timing and/or pattern of expression is consistent with this model (Ferris and Whitt, 1979; Hughes and Hughes, 1993; Force *et al.*, 1999). In addition, Force (1999) noted that, with both the engrailed genes, *engl* and *engl b*, in zebrafish and the ZAG*1* and ZMM2 genes in maize, the shared expression pattern of the duplicated genes matches the total expression pattern of single-copy genes in related organisms lacking the duplication. This observation could, however, be a result of either positive selection favoring specialization (Hughes, 1994) or loss of subfunctions through deleterious mutation accumulation (Force *et al.*, 1999; Lynch and Force, 2000).

III. DISCUSSION

Arguably, the duplication of genes, regulatory regions, and whole genomes has provided the most important raw material for evolution, for, without such duplications, it is likely that genomes would be limited to a handful of genes and organisms to a handful of components. Nevertheless, evolution has no foresight, and populations cannot evolve so as to maximize their future rate of adaptive evolution. In fact, the rate of gene and genome duplication is unlikely to be set at an optimal level for adaptive evolution. As with mutations in general (Sniegowski *et al.*, 2000), because gene duplications and genome duplications are more likely to reduce fitness than they are to increase fitness, individuals carrying genes that increase the rate of duplication ("duplicators") will typically have a lower fitness, being burdened by the disruptive effects of gene duplication. Even if a beneficial duplication does arise in an individual carrying a "duplicator," recombination will ensure that the beneficial duplicate spreads from the "duplicator" genotype to "nonduplicator" genotypes within the population. Thus there is an asymmetry in the effectiveness of selection. Deleterious duplications immediately decrease the average fitness of duplicator genotypes, whereas beneficial duplications only increase the average fitness of duplicator genotypes to the extent that they remain associated with the genes causing the increased rate of duplication. In sexual populations, this beneficial association decays rapidly and provides only a weak and ineffective advantage to the duplicator genes. (See Sniegowski *et al.*, 2000, for a review of the relevant theory on mutators, most of which may be applied directly to the case of duplicators.) Consequently, the most likely explanation for why gene and genome duplications occur is that it is too costly, in terms of time and energy, to eliminate these errors entirely.

The fact that gene and genome duplications are unlikely to have arisen to promote the evolvability of a population does not mean that they have not subsequently done so. Indeed, one could imagine that an organism early in evolution with only a handful of genes would have been more limited in evolutionary potential compared with any competitors that had undergone a spate of gene duplication. Consequently, it may be that organisms lacking duplications in their past have been more likely to go extinct or less likely to speciate over evolutionary time (although there is no clear evidence to this effect). This is, of course, a group selection argument. Although group selection is notoriously ineffective relative to individual selective forces acting on duplicators such as the ones described above, group selection can play an important role in determining which species have survived to comprise the biodiversity surrounding us. It remains a tantalizing possibility that the diversity and evolutionary success of ray-finned fishes and vertebrates, for example, were driven initially be genome-wide duplications. Yet the fact that there is no clear relationship between complexity (measured, for example, by the number of cell morphotypes; Valentine, 2000) and genome size, gene number, or proteome size suggests that we must be cautious in estimating the evolutionary importance of sheer numbers of genes. Complexity can and has evolved by other means, including the evolution of multidomain proteins, complex patterns of gene regulation, and alternative splicing.

Once a gene duplication appears, selection, drift, and mutation may each play a role in determining its ultimate fate. Although many theoretical analyses have focussed on the fate of a duplicate gene once it has already fixed within a population (with notable exceptions, e.g., Spofford, 1969; Ohta, 1987, 1988b, 1988c; Clark, 1994), evolutionary forces can have a major effect on the rate of spread and the probability of fixation of a new duplicate. Selection, for example, favoring a new gene duplicate improves the chance that the duplicate survives and rises in frequency. Thus, as we point out, duplicate genes that have fixed within a population will be relatively enriched for duplicates with a positive selective effect (Figure 16.1). Such fitness advantages may arise in a number of ways, including increasing the expression of underexpressed genes, allowing permanent heterozygosity when selection favors individuals with multiple alleles of a gene, and masking the deleterious effects of mutations. In an Appendix, we have constructed a model that examines the spread of a new gene duplication under various types of selection, with arbitrary linkage between the gene and its duplicate. We find that heterozygote advantage does indeed provide a large fitness advantage to a gene duplication. In fact, the strength of selection favoring a duplicated gene is comparable to the amount of selection favoring heterozygotes over homozygotes. Indeed, one reason why heterozygote advantage may be uncommon is that whenever this form of selection arises it leads, over evolutionary time, to the spread of gene duplicates such that every individual carries both alleles (Haldane, 1954).

In contrast, masking of deleterious mutations has only a very minor effect on the fitness of individuals bearing a duplicate gene. Even though it might be optimal to have a backup copy of a gene in case it mutates, selection is simply not powerful enough to cause the spread of gene duplicates to accomplish this benefit. With heterozygote advantage, a lower recombination rate (i.e., tighter linkage) between the gene and its duplicate promotes the spread of the duplication. With masking of deleterious mutations, a higher recombination rate (i.e., looser linkage) between the gene and its duplicate promotes the spread of the duplication. Given that duplications are much more strongly selected in the case of heterozygote advantage, these results suggest that tandem duplications are likely to have a stronger average benefit, making them more likely to establish when they occur compared to dispersed gene duplications. (See the Appendix for further details and discussion.)

Through selection or genetic drift, some gene duplications do fix within a population. However, they may still be lost or inactivated, especially if deleterious mutations accumulate as a result of the redundancy provided by gene copies. Although pseudogenes provide ample evidence for such inactivation, many duplicate genes have maintained their function over long periods of time, and still others have gained new functions. Maintenance of gene function is likely to occur in those cases where the effects of the duplicate gene are immediately advantageous. Divergence in gene function is likely to occur for genes that have multiple domains or multiple regulatory regions, both because selection might favor specialization of different gene copies on different functions and because the two gene copies might suffer deleterious mutations in different components, making both copies essential. Analyses of genomic sequences have clarified the role that different evolutionary processes have played in the origin, spread, maintenance, and diversification of duplicated genes (Lundin, 1993; Wolfe and Shields, 1997; Postlethwait *et al.*, 1998; Lynch and Conery, 2000; Vision *et al.*, 2000; Arabidopsis Genome Initiative, 2000). Undoubtedly, as more genomic sequence data become available, especially for closely related species, we will gain a clearer appreciation for the role that gene duplication has played in the evolution of life.

IV. APPENDIX: MODELING THE INVASION OF A DUPLICATE GENE

Most models of gene duplication assume that a gene duplicate has already fixed within the population or that duplicates are continuously produced (see Walsh, 1995). The two notable exceptions are the studies by Spofford (1969) and Clark (1994). Spofford focused on the spread of a chromosome carrying two different alleles (A and a) when there is a fitness advantage to individuals carrying both

alleles, finding graphically that spread of the duplicate always occurred. Clark, on the other hand, modeled the situation in which deleterious mutations occur at both the original and duplicate genes. Assuming that these genes are completely linked and that only individuals lacking an A allele suffer a fitness loss, he found that the duplicate gene would spread only if it had a direct advantage. Here, we generalize these results by examining the initial rate of spread of a duplicate gene under a broader array of conditions. Specifically, we calculate the magnitude of selection for or against the gene duplicate under more general fitness regimes and with an arbitrary linkage relationship between the duplicate genes. To simplify matters, we ignore the ongoing production of gene duplicates (as was examined by Clark, 1994, and by Ohta, 1987, 1988b, 1988c) and gene conversion, both of which may play an important role in the establishment of a gene duplicate.

In our model, we consider an original locus (i) and its duplicate (j), separated by distance r. We denote a chromosome lacking a duplication by i- and one carrying the duplication by ij. Duplications on different chromosomes may be examined by setting r to $\frac{1}{2}$. At each gene there are two alleles A and a. Thus, with the duplication, there are four haplotypes to consider (AA, Aa, aA, and aa), whose frequencies are x_1, x_2, x_3, and x_4. Without the duplication, there are two additional haplotypes to consider (denoted as A- and a-), whose frequencies are x_5 and x_6. The fitness of a diploid individual carrying haplotypes k and l is denoted W_{kl}. At each gene, mutations are allowed to occur between alleles A and a at rate μ. To leading order in the mutation rate, mutations affect the invasion criterion only when there is no direct selection on the gene duplication and when the population is initially near a mutation-selection balance (Yong, 1998). Therefore we only discuss the effects of mutations in this case.

A. General diploid model

We model a diploid population that is at equilibrium when a gene duplication occurs. At this equilibrium, x_1, x_2, x_3, and x_4 equal zero, $x_5 = \hat{x}_5$, $x_6 = \hat{x}_6$, $\hat{x}_5 + \hat{x}_6 = 1$, and the mean fitness of the population is $\hat{W}$. A gene duplication then arises at low frequency. While the duplicate is rare, haplotypes bearing the duplication will change in frequency from one generation (x) to the next (x') according to the recursions:

$$\begin{aligned} x_1' &= (1-\mu)^2 x_1^* \\ x_2' &= (1-\mu)x_2^* + \mu(1-\mu)x_1^* \\ x_3' &= (1-\mu)x_3^* + \mu(1-\mu)x_1^* \\ x_4' &= x_4^* + \mu^2 x_1^* + \mu x_2^* + \mu x_3^* \end{aligned} \tag{A1}$$

where

$$
\begin{aligned}
x_1^* &\approx \frac{x_1\hat{x}_5W_{15} + r\,x_3\hat{x}_5W_{35} + (1-r)x_1\hat{x}_6W_{16}}{\hat{W}} \\
x_2^* &\approx \frac{x_2\hat{x}_5W_{25} + r\,x_4\hat{x}_5W_{45} + (1-r)x_2\hat{x}_6W_{26}}{\hat{W}} \\
x_3^* &\approx \frac{x_3\hat{x}_6W_{36} + r\,x_1\hat{x}_6W_{16} + (1-r)x_3\hat{x}_5W_{35}}{\hat{W}} \\
x_4^* &\approx \frac{x_4\hat{x}_6W_{46} + r\,x_2\hat{x}_6W_{26} + (1-r)x_4\hat{x}_5W_{45}}{\hat{W}}
\end{aligned}
\tag{A2}
$$

Technically, equation (A1) is written assuming that mutations generate only a alleles from A alleles. However, if mutations occur rarely and are deleterious, the a allele will exist at low frequency within a population, and reverse mutations to it will be extremely rare. Therefore, back-mutations have little effect on the dynamics and can be safely ignored. Equations (A1) are linear functions of the frequencies of the duplicate gene because genotypes involving two duplication haplotypes are assumed to be exceedingly rare when the duplicate first appears and are ignored. To determine whether the duplicate gene spreads, we perform a local stability analysis of these equations (Bretscher, 1997). This involves finding the leading eigenvalue, λ_L, of the matrix form of equation (A1), which is known as a local stability matrix. $(\lambda_L - 1)$ corresponds, approximately, to the selection coefficient that acts on the gene duplicate while it is rare (Otto and Bourguet, 1999). Consequently, the duplicate will initially increase in frequency at an exponential rate whenever $\lambda_L > 1$ (or decrease if $\lambda_L < 1$). We shall first consider the rate of spread of the duplicate in a model with heterozygous advantage and then consider its spread when there are recurrent deleterious mutations.

B. Heterozygote advantage

The first case that we shall consider is heterozygote advantage. We assume that alleles A and a perform different functions such that the fitness of an individual is higher when both functions are performed. To simplify matters, we assume that all individuals carrying both A and a alleles (regardless of the number of them) have a fitness of $1(= W_{56} = W_{25} = W_{35} = W_{45} = W_{16} = W_{26} = W_{36})$, that individuals with only A alleles have a reduced fitness of $1 - s(= W_{55} = W_{15})$, and that individuals with only a alleles have a reduced fitness of $1 - t(= W_{66} = W_{46})$. This model differs from that examined by Spofford (1969), who focused on the case of a dimeric enzyme taking into account the probability that each type of dimer would be produced. In addition, the following derivation estimates the effective selection coefficient acting on a new duplicate, whereas Spofford focused on a numerical analysis of the dynamics of the duplicate gene.

Before the appearance of the duplication, the frequency of allele A approaches $\hat{x}_5 = t/(s+t)$, at which point the mean fitness of the population is $\hat{W} = 1 - st/(s+t)$ (Crow and Kimura, 1970). Without loss of generality, we label the alleles such that $s \leq t$, so that the frequency of allele A is greater than or equal to $\frac{1}{2}$. Under these assumptions, it can be shown that the leading eigenvalue describing the initial rate of spread of a gene duplicate is equal to:

$$\lambda_L = \frac{1}{\hat{W}} - \frac{1 - \hat{W} + r}{2\hat{W}} \left[1 - \sqrt{1 - \frac{4r\hat{x}_6(1-\hat{W})}{(1-\hat{W}+r)^2}} \right] \tag{A3}$$

Under our assumptions that s, t, $\hat{x}_5$, $\hat{x}_6 > 0$, this leading eigenvalue is strictly greater than one. Thus, there will always be selection favoring the spread of a duplication. When the duplicate genes are tightly linked ($r \approx 0$), the leading eigenvalue becomes $1/\hat{W}$, which is greater than one under our assumptions. In this case, there is a slight caveat: if the first haplotype to appear is either AA or aa, the duplication will not be positively selected until either the Aa or aA haplotype is produced by mutation, conversion, or recombination. Increasing the recombination rate above zero always reduces the leading eigenvalue and hence slows the initial spread of the duplicate gene. When selection is weak ($s, t \ll 1$) and when the genes are not very tightly linked ($r > 0$), the eigenvalue is approximately:

$$\lambda_L \approx 1 + \hat{x}_5(1 - \hat{W}) = 1 + \frac{st^2}{(s+t)^2} \tag{A4}$$

Equation (A4) indicates that the indirect or effective selection acting on the duplicate has the same order of magnitude as selection acting directly on the A and a alleles. Furthermore, selection for the duplicate is strongest when the two functions are approximately of equal benefit ($s \approx t$). Comparing the strength of selection on the duplicate ($\lambda_L - 1$), the duplicate experiences $(s+t)/t$ times the amount of selection when linkage is tight than when linkage is loose, assuming weak selection. This reaches a maximum of a twofold difference when $s \approx t$, indicating that a tandem duplication carrying both A and a alleles is twice as likely to fix within a population (replacing $\lambda_L - 1$ for the selection coefficient in equation 1) and will spread twice as quickly compared to an unlinked gene duplication. The advantage of tandem duplications is that the most fit chromosomes, i.e., ones bearing both A and a alleles, are unlikely to be broken apart by recombination. This may partially explain why tandem gene duplications are common in many gene families, although another obvious explanation is that the frequency with which duplications appear in tandem is higher as a result of unequal crossing over.

C. Mutation-selection balance

We next examine the fate of the gene duplication when the population is at mutation-selection balance. In the absence of the duplication, let the fitness of AA, Aa, and aa genotypes equal $W_{55} = 1$, $W_{56} = 1 - hs$, and $W_{66} = 1 - s$, respectively, where s measures the fitness cost of the mutation and h measures the degree of dominance of the a allele. We first assume that mutations are not fully recessive ($h > 0$). In large populations, the mutant allele (a) will reach a steady-state frequency of $\hat{x}_6 \approx \mu/(hs)$, at which point the population mean fitness is $\hat{W} \approx 1 - 2\mu$ (Crow and Kimura, 1970). These approximations are to leading order in the mutation rate and ignore terms such as μ^2. When the duplication first appears, we assume that it experiences no direct selection, such that the fitness of AA/A-individuals is also $1 (= W_{15})$. Individuals carrying one mutant allele (AA/a- or Aa/A- or aA/A-) are assumed to have a fitness of $1 - ks (= W_{25} = W_{35})$, where k is the dominance coefficient of the a allele when paired with two A alleles. It can be shown that the fitness of genotypes bearing more than one deleterious mutation does not affect the spread of the gene duplication since these genotypes are much less common. Making these substitutions as well as $\hat{x}_5 = 1 - \hat{x}_6$ in equation (A1), we obtain a local stability matrix that is solely a function of the mutation rate and selection parameters. The leading eigenvalue of this matrix can be shown to equal

$$\lambda_L = 1 - \mu\left\{\frac{k}{h} - \frac{(h-k)\,r\,(1-ks)}{h[r + ks(1-r)]}\right\} + O(\mu^2) \qquad \text{(A5)}$$

Assuming that the a allele is always deleterious, the term k/h is positive. Hence, for tandemly duplicated genes ($r \approx 0$), the leading eigenvalue is always less than one. Counterintuitively, this means that the duplication will spread at a slower rate than a purely neutral allele even if the duplication provides extra protection against the deleterious effects of mutations (i.e., $k < h$). The reason that tandem duplications are selected against is that mutations occur twice as often on chromosomes with two genes and are not strongly selected against when $k < h$. Consequently, deleterious mutant alleles accumulate on the chromosome bearing the duplication, generating indirect selection against the duplication. Note, however, that the strength of selection against the duplicate is proportional to the mutation rate, μ. Therefore, this effect is so slight that it will not completely prevent the fixation of any particular gene duplicate, but it will reduce the rate at which duplicate genes fix within large populations.

It can be shown that the leading eigenvalue, (A5), is always greater for gene duplicates that arise at a distance from the original gene (i.e., $\partial\lambda_L/\partial r > 0$).

Nevertheless, only when recombination is high enough,

$$r > \frac{k^2 s}{(h - 2k)(1 - ks)} \tag{A6}$$

and the masking advantage of the new duplicate is strong enough,

$$k < \frac{h}{2 + hs} \tag{A7}$$

will the gene duplication be positively selected ($\lambda_L - 1 > 0$). For weak selection and unlinked gene duplicates, these conditions require that the fitness reduction observed in AAa individuals be less than half that observed in Aa individuals ($k < h/2$). The reason that distant gene duplicates fare better is that recombination shuffles some of the deleterious mutant alleles that accumulate faster in individuals bearing duplicated genes to chromosomes lacking the gene duplication. This reduces the frequency of deleterious mutant alleles associated with the gene duplication. Even when the gene duplicate is favored, the strength of this indirect selective advantage is always less than twice the mutation rate, and hence masking deleterious mutations will only slightly increase the chance that loosely linked duplicates spread to fixation.

The results obtained in this section are similar to those obtained in models of ploidy evolution, where it has been shown that selection only favors the expansion of the diploid phase of a sexual organism (with "duplicates" of every gene) when recombination rates are sufficiently high and masking is sufficiently strong (Otto and Goldstein, 1992; Jenkins, 1995). The main qualitative difference is that duplication of the entire genome allows the effects to be greatly amplified, relative to the case of a single gene duplicate (Jenkins, 1995).

D. Mutation-selection balance with completely recessive mutations

In the above, we assumed that deleterious mutations have some deleterious effects even in heterozygous individuals. Here, we briefly review the results obtained when masking is complete ($h = k = 0$), such that only aa and aa/a- individuals have a reduced fitness of $1 - s$. In this case, before the appearance of the duplicate, $\hat{x}_6 \approx \sqrt{\mu/s}$ and $\hat{W} \approx 1 - \mu$ at the equilibrium between mutation and selection (Crow and Kimura, 1970). Repeating the local stability analysis, we find that the leading eigenvalue is always one, regardless of the recombination rate. This is the same result obtained by Clark (1994) for tandem duplications with lethal homozygous effects ($r = 0, s = 1$). Thus, the additional protection against deleterious mutations provided by a gene duplication makes no difference to the spread of the duplication when deleterious mutations are already fully masked in Aa heterozygotes lacking the duplicate.

E. Further results

Although the above analyses assume that the duplication is not directly selected for or against, it is straightforward to include such selection. Let s_D equal the direct selection coefficient (i.e., the change in fitness due to having a gene duplicate regardless of which alleles are carried). Similarly, let s_I equal the indirect selection that arises because carrying a gene duplicate changes the alleles that an individual is likely to carry, as measured in the above sections [i.e., $s_I = \lambda_L - 1$ from equation (A3) or (A5)]. Assuming that selection is relatively weak (S_D, $S_I \ll 1$), the total strength of selection acting on the gene duplicate while rare is simply the sum of these two effects (Yong, 1998; Otto and Bourguet, 1999). Thus, any amount of direct selection is likely to overwhelm the indirect selection that arises from deleterious mutations alone (from A5). In contrast, when there is heterozygote advantage, the strength of indirect selection (from A3) can be quite strong and may control the fate of a new gene duplicate.

Finally, although we have focused on the dynamics of gene duplications in diploid organisms, in his thesis, Yong (1998) also examined the dynamics of gene duplication in haploid organisms. Each generation, the haploid organisms were assumed to mate randomly to form diploid individuals, followed by meiosis to form daughter haploid organisms. One surprising difference in the results of the haploid and diploid models was that gene duplications in haploids never gain an advantage from masking deleterious mutations. In other words, in a haploid mutation-selection balance model, the leading eigenvalue (equivalent to A5) was always less than one, even though individuals carrying a gene duplication and a mutant allele (i.e., *Aa* individuals) were more fit than individuals carrying only a mutant allele (i.e., *a-* individuals). We speculate that this difference is caused by the fact that mutations accumulate more rapidly in gene duplicates in haploids than in diploids. Consider individuals carrying a new gene duplicate. If they are diploid, a new mutation occurs in the gene duplicate only one-third of the time; two-thirds of the time it will occur in the original gene. Because the gene duplication provides more opportunity for masking the effects of this new mutation, this mutation is sheltered from selection and has a higher chance of being passed to the offspring generation. Nevertheless, the gene duplicate is mutation free two-thirds of the time and, with loose linkage, has a good chance of segregating away from the mutation. In haploids, however, a new mutation that is sheltered by a gene duplication has a 50:50 chance of having occurred in the duplicate gene itself, in which case the duplicate and the new mutation are inextricably bound. Thus, even with free recombination, there is less opportunity for new gene duplicates to rid themselves of the higher frequency of mutant alleles that they shelter from selection. Hence, in haploids, the masking advantage of a gene duplication is always outweighed by the disadvantage that is incurred because mutations are sheltered from selection and reach a higher frequency at the duplicated gene.

Acknowledgments

The models described in the Appendix were inspired by the work of Andrew Clark and benefited greatly from discussions with him. Further mathematical details may be found in the Honour's thesis of PY, available from the Department of Zoology, University of British Columbia. We are grateful to Jay Dunlap, Thomas Lenormand, Toby Johnson, and Michael Whitlock for helpful discussions and comments on the manuscript. Funding to SPO was provided by a grant from the Natural Sciences and Engineering Research Council (Canada) and by a poste-rouge from the Centre National de la Recherche Scientifique (France).

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., and Galle, R. F. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195.

Ahn, S., and Tanksley, S. D. (1993). Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. USA* **90,** 7980–7984.

Al-Kaff, N. S., Covey, S. N., Kreike, M. M., Page, A. M., Pinder, R., and Dale, P. J. (1998). Transcriptional and posttranscriptional plant gene silencing in response to a pathogen. *Science* **279,** 2113–2115.

Amores, A., Force, A., Yan, Y. L., Joly, J. S., Amemiya, C. T., Fritz, A., Ho, R. K., Langeland, J., Prince, V., and Wang, Y. L. (1998). Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282,** 1711–1714.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.

Bailey, G. S., Poulter, R. T., and Stockwell, P. A. (1978). Gene duplication in tetraploid fish: Model for gene silencing at unlinked duplicated loci. *Proc. Natl. Acad. Sci. USA* **75,** 5575–5579.

Baldo, A. M., and McClure, M. A. (1999). Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts. *J. Virol.* **73,** 7710–7721.

Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., and Mayhew, G. F. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474.

Bretscher, O. (1997). "Linear Algebra with Applications." Prentice Hall, Upper Saddle River, NJ.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282,** 2012–2018.

Charlesworth, B., and Langley, C. H. (1991). Population genetics of transposable elements in *Drosophila*. *In* "Evolution at the Molecular Level" (R. K. Selander, A. G. Clark, and T. S. Whittam, eds.), pp. 222–247. Sinauer, Sunderland, MA.

Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., and Smith, T. (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282,** 2022–2028.

Clark, A. G. (1994). Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. USA* **91,** 2950–2954.

Cooke, J., Nowak, M. A., Boerlijst, M., and Maynard-Smith, J. (1997). Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* **13,** 360–364.

Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., and Quetier, F. (2000). Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238.

Crow, J. F., and Kimura, M. (1970). "An Introduction to Population Genetics Theory," pp. xiv, 591. Harper & Row, New York.

Deng, H. W., and Lynch, M. (1997). Inbreeding depression and inferred deleterious-mutation parameters in *Daphnia*. *Genetics* **147,** 147–155.

Duret, L. (2000). tRNA gene number and codon usage in the C. *elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16,** 287–289.

Elena, S. F., and Lenski, R. E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390,** 395–398.

Ewing, B., and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25,** 232–234.

Ferris, S. D., and Whitt, G. S. (1977). Loss of duplicate gene expression after polyploidisation. *Nature* **265,** 258–260.

Ferris, S. D., and Whitt, G. S. (1979). Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* **12,** 267–317.

Fisher, R. A. (1930). "The Genetical Theory of Natural Selection." Oxford University Press, Oxford, U.K.

Fisher, R. A. (1935). The sheltering of lethals. *Am. Naturalist* **69,** 446–455.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A, Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545.

Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S., and Fink, G. R. (1999). Ploidy regulation of gene expression. *Science* **285,** 251–254.

Gaut, B. S., and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA* **94,** 6809–6814.

Gelbart, W. M., and Chovnick, A. (1979). Spontaneous unequal exchange in the rosy region of *Drosophila melanogaster*. *Genetics* **92,** 849–859.

Gibson, T. J., and Spring, J. (2000). Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans.* **28,** 259–264.

Goldblatt, P. (1980). Polyploidy in angiosperms: Monocotyledons. *In* "Polyploidy: Biological Relevance" (W. H. Lewis, ed.), pp. 219–239. Plenum, New York.

Grant, V. (1963). "The Origin of Adaptations." Columbia University Press, New York.

Grant, V. (1981). "Plant Speciation." Columbia University Press, New York.

Guillemaud, T., Raymond, M., Tsagkarakou, A., Bernard, C., Rochard, P., and Pasteur, N. (1999). Quantitative variation and selection of esterase gene amplification in *Culex pipiens*. *Heredity* **83,** 87–99.

Haldane, J. B. S. (1933). The part played by recurrent mutation in evolution. *Am. Naturalist* **67,** 5–19.

Haldane, J. B. S. (1954). "The Biochemistry of Genetics." George Allen & Unwin, London.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B. Biol. Sci.* **256,** 119–124.

Hughes, A. L., and Nei, M. (1989). Evolution of the major histocompatibility complex: Independent origin of nonclassical class I genes in different groups of mammals. *Mol. Biol. Evol.* **6,** 559–579.

Hughes, M. K., and Hughes, A. L. (1993). Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10,** 1360–1369.

Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C. (1999). Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286,** 2165–2169.

Jenkins, C. D., and Kirkpatrick, M. (1995). Deleterious mutations and the evolution of genetic life cycles. *Evolution* **49,** 512–520.

Kimura, M. (1983). "The Neutral Theory of Molecular Evolution." Cambridge University Press, Cambridge, U.K.

Leipoldt, M. (1983). Towards an understanding of the molecular mechanisms regulating gene expression during diploidization in phylogenetically polyploid lower vertebrates. *Hum. Genet.* **65,** 11–18.

Lenormand, T., Guillemaud, T., Bourguet, D., and Raymond, M. (1998). Appearance and sweep of a gene duplication: Adaptive response and potential for new functions in the mosquito *Culex pipiens*. *Evolution* **52,** 1705–1712.

Lewin, B. (1994). "Genes V." Oxford University Press, Oxford, U.K.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet* **25,** 239–240.

Löve, A., Löve, D., and Pichi Sermolli, R. E. G. (1977). "Cytotaxonomical Atlas of the Pteridophyta." J. Cramer, Vaduz, Germany.

Lundin, L. G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16,** 1–19.

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154,** 459–473.

Maeda, N., and Smithies, O. (1986). The evolution of multigene families: Human haptoglobin genes. *Annu. Rev. Genet.* **20,** 81–108.

Maroni, G., Wise, J., Young, J. E., and Otto, E. (1987). Metallothionein gene duplications and metal tolerance in natural populations of *Drosophila melanogaster*. *Genetics* **117,** 739–744.

Masterson, J. (1994). Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* **264,** 421–423.

Maynard Smith, J. (1998). "Evolutionary Genetics." Oxford University Press, Oxford, U.K.

McFadden, D., Kwong, L., Yam, I., and Langlois, S. (1993). Parental origin of triploidy in human fetuses: Evidence for genomic imprinting. *Hum. Genet.* **92,** 465–469.

Mittelsten Scheid, O., Jakovleva, L., Afsar, K., Maluszynska, J., and Paszkowski, J. (1996). A change in ploidy can modify epigenetic silencing. *Proc. Natl. Acad. Sci. USA* **93,** 7114–7119.

Mittelsten Scheid, O., Paszkowski, J., and Potrykus, I. (1991). Reversible inactivation of a transgene in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **228,** 104–112.

Mouchès, C., Pasteur, N., Bergé, J. B., Hyrien, O., Raymond, M., de Saint Vincent, B. R., de Silvestri, M., and Georghiou, G. P. (1986). Amplification of an esterase gene is responsible for insecticide resistance in a California *Culex* mosquito. *Science* **233,** 778–780.

Mukai, T., Chigusa, S. I., Mettler, L. E., and Crow, J. F. (1972). Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. *Genetics* **72,** 335–355.

Nadeau, J. H., and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147,** 1259–1266.

Nowak, M. A., Boerlijst, M. C., Cooke, J., and Smith, J. M. (1997). Evolution of genetic redundancy. *Nature* **388,** 167–171.

Nur, U. (1980). Evolution of unusual chromosome systems in scale insects (Coccoidea: Homoptera). *In* "Insect Cytogenetics" (R. L. Blackman, G. M. Hewitt, and M. Ashburner, eds.), pp. 97–117. Blackwell Scientific, London.

Ohno, S. (1970). "Evolution by Gene Duplication." George Allen & Unwin, London.

Ohta, T. (1987). Simulating evolution by gene duplication. *Genetics* **115,** 207–213.

Ohta, T. (1988a). Evolution by gene duplication and compensatory advantageous mutations. *Genetics* **120,** 841–847.

Ohta, T. (1988b). Further simulation studies on evolution by gene duplication. *Evolution* **42,** 375–386.

Ohta, T. (1988c). Time for acquiring a new gene by duplication. *Proc. Natl. Acad. Sci. USA* **85,** 3509–3512.

Otto, S. P., and Bourguet, D. (1999). Balanced polymorphisms and the evolution of dominance. *Am. Naturalist* **153,** 561–574.

Otto, S. P., and Goldstein, D. B. (1992). Recombination and the evolution of diploidy. *Genetics* **131,** 745–751.

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* **34,** 401–437.

Pál, C., and Hurst, L. D. (2000). The evolution of gene number: Are heritable and non-heritable errors equally important? *Heredity* **84,** 393–400.

Postlethwait, J. H., Yan, Y. L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., and Gong, Z. (1998). Vertebrate genome evolution and the zebrafish gene map. *Nature Genet.* **18,** 345–349.

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* **29,** 467–501.

Ratcliff, F., Harrison, B. D., and Baulcombe, D. C. (1997). A similarity between viral defense and gene silencing in plants. *Science* **276,** 1558–1560.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., and Fleischmann, W. (2000). Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215.

Selker, E. U. (1997). Epigenetic phenomena in filamentous fungi: Useful paradigms or repeat-induced confusion? *Trends Genet.* **13,** 296–308.

Selker, E. U. (1999). Gene silencing: Repeats that count. *Cell* **97,** 157–160.

Seoighe, C., and Wolfe, K. H. (1999). Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* **2,** 548–554.

Shaw, R. G., Byers, D. L., and Darmo, E. (2000). Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics* **155,** 369–378.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407,** 81–86.

Simmons, M. J., and Crow, J. F. (1977). Mutations affecting fitness in *Drosophila* populations. *Annu. Rev. Genet.* **11,** 49–78.

Skrabanek, L., and Wolfe, K. H. (1998). Eukaryote genome duplication—Where's the evidence? *Curr. Opin. Genet. Dev.* **8,** 694–700.

Sniegowski, P. D., Gerrish, P. J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: Separating causes from consequences. *Bioessays* **22,** 1057–1066.

Soltis, D. E., and Soltis, P. S. (1993). Molecular data and the dynamic nature of polyploidy. *Crit. Rev. Plant Sci.* **12,** 243–273.

Soltis, D. E., and Soltis, P. S. (1999). Polyploidy: Recurrent formation and genome evolution. *Trends Ecol. Evol.* **14,** 348–352.

Spofford, J. B. (1969). Heterosis and the evolution of duplications. *Am. Naturalist* **103.**

Stebbins, G. L. (1938). Cytological characteristics associated with the different growth habits in the dicotyledons. *Am. J. Bot.* **25,** 189–198.

Stebbins, G. L. (1980). Polyploidy in plants: Unsolved problems and prospects. *In* "Polyploidy: Biological Relevance" (W. H. Lewis, ed.), pp. 495–520. Plenum, New York..

Takahata, N., and Maruyama, T. (1979). Polymorpism and loss of duplicate gene expression: A theoretical study with application to tetraploid fish. *Proc. Natl. Acad. Sci. USA* **76,** 4521–4525.

Tekaia, F., and Dujon, B. (1999). Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J. Mol. Evol.* **49,** 591–600.

Uchida, I. A., and Freeman, V. C. (1985). Triploidy and chromosomes. *Am. J. Obstet. Gynecol.* **151,** 65–69.

Valentine, J. W. (2000). Two genomic paths to the evolution of complexity in bodyplans. *Paleobiology* **26,** 513–519.

Vassilieva, L. L., Hook, A. M., and Lynch, M. (2000). The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. *Evolution* **54,** 1234–1246.

Veiko, N. N., Lyapunova, N. A., Bogush, A. I., Tsvetkova, T. G., and Gromova, E. V. (1996). Ribosomal gene number in individual human genomes—data from comparative molecular and cytogenetic analysis. *Mol. Biol.* **30,** 641–647.

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in *Arabidopsis*. *Science* **290,** 2114–2117.

Vrijenhoek, R., Dawley, R., Cole, C., and Bogart, J. (1989). A list of known unisexual vertebrates. *In* "Evolution and Cytology of Unisexual Vertebrates" (R. Dawley, and J. Bogart, eds.), pp. 19–23. The University of the State of New York, New York.

Walsh, J. B. (1987). Sequence-dependent gene conversion: Can duplicated genes diverge fast enough to escape conversion? *Genetics* **117,** 543–557.

Walsh, J. B. (1995). How often do duplicated genes evolve new functions? *Genetics* **139,** 421–428.

Wang, J. Y., McCommas, S., and Syvanen, M. (1991). Molecular cloning of a glutathione S-transferase overproduced in an insecticide-resistant strain of the housefly (*Musca domestica*). *Mol. Gen. Genet.* **227,** 260–266.

Watterson, G. A. (1983). On the time for gene silencing at duplicate loci. *Genetics* **105,** 745–766.

Werth, C., and Windham, M. (1991). A model for divergent, allopatric speciation of polyploidy pteridophytes resulting from silencing of duplicate-gene expression. *Am. Naturalist* **137,** 515–526.

Wolf, S., Sharpe, L. T., Schmidt, H. J. A., Knau, H., Weitz, S., Kioschis, P., Poustka, A., Zrenner, E., Lichter, P., and Wissinger, B. (1999). Direct visual resolution of gene copy number in the human photopigment gene array. *Invest. Ophthalmol. Visual Sci.* **40,** 1585–1589.

Wolfe, K. H., and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713.

Wolffe, A. P., and Matzke, M. A. (1999). Epigenetics: Regulation through repression. *Science* **286,** 481–486.

Wu, C. T., and Morris, J. R. (1999). Transvection and other homology effects. *Curr. Opin. Genet. Dev.* **9,** 237–246.

Yong, P. (1998). Theoretical population genetic model of the invasion of an initial duplication, Honour's thesis, Department of Zoology, University of British Colombia, Vancouver.