The Probability of Duplicate Gene Preservation by Subfunctionalization

Michael Lynch and Allan Force

Department of Biology, University of Oregon, Eugene, Oregon 97403

Manuscript received June 19, 1999
Accepted for publication September 15, 1999

ABSTRACT

It has often been argued that gene-duplication events are most commonly followed by a mutational event that silences one member of the pair, while on rare occasions both members of the pair are preserved as one acquires a mutation with a beneficial function and the other retains the original function. However, empirical evidence from genome duplication events suggests that gene duplicates are preserved in genomes far more commonly and for periods far in excess of the expectations under this model, and whereas some gene duplicates clearly evolve new functions, there is little evidence that this is the most common mechanism of duplicate-gene preservation. An alternative hypothesis is that gene duplicates are frequently preserved by subfunctionalization, whereby both members of a pair experience degenerative mutations that reduce their joint levels and patterns of activity to that of the single ancestral gene. We consider the ways in which the probability of duplicate-gene preservation by such complementary mutations is modified by aspects of gene structure, degree of linkage, mutation rates and effects, and population size. Even if most mutations cause complete loss-of-subfunction, the probability of duplicate-gene preservation can be appreciable if the long-term effective population size is on the order of 10^5 or smaller, especially if there are more than two independently mutable subfunctions per locus. Even a moderate incidence of partial loss-of-function mutations greatly elevates the probability of preservation. The model proposed herein leads to quantitative predictions that are consistent with observations on the frequency of long-term duplicate gene preservation and with observations that indicate that a common fate of the members of duplicate-gene pairs is the partitioning of tissue-specific patterns of expression of the ancestral gene.

DUPLICATE genes arise frequently in eukaryotic genomes, either via local events that generate tandem duplications, larger-scale events that duplicate chromosomal regions or entire chromosomes, or genome-wide events that result in complete genome duplication (polyploidization). Because gene duplicates are believed to be initially redundant in function, it is commonly thought that one member of the pair will usually become silenced by degenerative mutation. Such nonfunctionalization is expected to occur within a few million generations because the rate of mutation to null alleles is on the order of 10^{-6} per generation, while the incidence of mutations to novel and beneficial functions is much lower. Although this classical model has been subject to substantial mathematical analysis (Haldane 1933; Fisher 1935; Nei and Roychoudhury 1973; Christiansen and Frydenberg 1977; Bailey *et al.* 1978; Kimura and King 1979; Takahata and Maruyama 1979; Li 1980; Watterson 1983; Walsh 1995), it does not easily accommodate the existing data. It is now known that most eukaryotic genomes harbor large numbers of functional gene duplicates, many of which originated tens to hundreds of millions of years ago (Allendorf *et al.* 1975; Ferris and Whitt 1979; Graf and Kobel 1991; Lundin 1993; Sidow 1996; Brookfield 1997; Nadeau and Sankoff 1997; Postlethwait *et al.* 1998; Wendel 1999). The high degree of duplicate-gene preservation observed for genome duplication events (commonly on the order of 20–50%) for such long periods of time suggests that some type of positive selection must be offsetting the high rate of production of null alleles, and this is supported by frequent observations of rates of accumulation of expressed mutations in both members of a pair that are less than the neutral expectation (Li 1985; Hughes and Hughes 1993; Ramos-Onsins and Aguade 1998).

Under the classical model for the evolution of gene duplicates, the only mechanism by which members of a pair can permanently escape mutational decay is neofunctionalization, whereby one copy acquires a new beneficial function with the other retaining the original function (Ohno 1970; Ohta 1988; Walsh 1995; Nowak *et al.* 1997). However, other mechanisms for the preservation of duplicate genes can be envisioned. For example, instances may exist in which there is positive selection for the maintenance of multiple copies of genes (Ohta 1987; Clark 1994; Nowak *et al.* 1997; Wagner 1999). We have recently suggested an alternative mechanism by which duplicate genes may be commonly preserved. One limitation of the classical model for the evolution of gene duplicates is the implicit assumption that loss-of-function mutations simultaneously eliminate

Corresponding author: Michael Lynch, Department of Biology, University of Oregon, Eugene, OR 97403.
E-mail: mlynch@oregon.uoregon.edu

all aspects of gene expression. The duplication/degeneration/complementation (DDC) model (Force *et al.* 1999) derives from the fact that many genes, particularly those involved in development, have multiple, independently mutable subfunctions with respect to timing and tissue specificity of expression. With this more general view of gene structure, a plausible and parsimonious explanation for the long-term preservation of gene duplicates is the loss of different ancestral subfunctions by the two descendant members of the pair. Subfunctionalization is defined as the fixation of complementary loss-of-function alleles that results in the joint preservation of duplicate loci. For example, a gene that is originally expressed in two tissues may diverge into two copies, each being expressed uniquely in one of the two tissues. Provided the different subfunctions are essential for survival and/or reproduction, once such a partitioning of expression pattern has become fixed in a population, the two copies will be maintained indefinitely by natural selection. A unique feature of the subfunctionalization model for the evolution of duplicate genes is that gene preservation is entirely a consequence of degenerative mutations. Beneficial mutations need not be invoked. Stoltzfus (1999) has also suggested that partial loss-of-function mutations can lead to the preservation of duplicate genes with only a single function.

Numerous examples now exist for the presence of independently mutable regulatory sequences associated with developmental genes. Consider, for example, the *bmp5* gene in the mouse. In a study of 34 induced mutations at this locus, DiLeone *et al.* (1998) found seven alleles that exhibited no changes in the coding region of the gene. Studies of the tissue-specific expression patterns of these alleles revealed a large number of *cis*-acting regulatory elements, each driving expression at specific locations in the skeleton and other tissues, *e.g.*, the top of the sternum, genital tubercles, thyroid cartilage, intestine, and lungs. Many of the regulatory elements appeared to be located >270 kb from the transcription initiation site for the locus. For other well-documented examples of genes with modular structure for regulatory sequences, see Huang *et al.* (1993), Jack and DeLotto (1995), Slusarski *et al.* (1995), Kirchhamer *et al.* (1996), Gerhart and Kirschner (1997), and Arnone and Davidson (1997).

The idea that the differences in expression patterns of gene duplicates are often a consequence of evolutionary partitioning of the expression domains of ancestral genes, rather than reflecting the origin of new gene functions, is motivated by observations of tissue-specific patterns of expression of duplicate allozymes in tetraploid lineages of fish (Allendorf *et al.* 1975; Ferris and Whitt 1979). More recent studies of developmental genes have provided additional evidence for the preservation of duplicate genes by complementary loss of subfunctions. For example, we reported on the expression domains of two duplicate *engrailed* genes in zebrafish (Force *et al.* 1999). These two genes originated after the divergence of ray-finned fishes and tetrapods, and one is expressed in the pectoral appendage bud, while the other is expressed in hindbrain/spinal cord. In contrast, the single orthologous copy that is present in both mouse and chicken is expressed in both regions. Likewise, two *Notch* duplicates exist in zebrafish, one of which is expressed in presomitic mesoderm and the other in endocardial cells, whereas the single orthologous copy in mouse is expressed in both tissues (Westin and Lardelli 1997). A remarkably similar pattern is seen in the two zebrafish *Pax6* genes, which have unique expression patterns that sum to the total expression pattern for the single copy of *Pax6* present in birds and mammals (Normes *et al.* 1998). The modular nature of the tissue-specific regulatory regions of *Pax6* has been verified at the molecular level in mammals (Kammandel *et al.* 1999; Xu *et al.* 1999). The most parsimonious explanation for all of these observations is that subsequent to a complete genome duplication in the lineage containing the zebrafish (Amores *et al.* 1998; Postlethwait *et al.* 1998), the two members of each gene pair partitioned up the expression patterns of the ancestral gene, which remain as a single copy in the tetrapod lineage. Given that very few other attempts have been made to understand the evolution of gene expression patterns in a comparative phylogenetic framework, the high incidence of these types of observations [see Force *et al.* (1999) for other examples] suggests that loss-of-subfunction mutations are a common determinant of the fate of gene duplicates.

Although the DDC process is based entirely on degenerative mutations, there are at least three ways in which it may play a significant role in creative evolutionary processes. First, by stabilizing duplicate genes in the genome, the DDC process extends the time period during which genes are exposed to natural selection, thereby enhancing the chance that rare beneficial mutations to novel functions may arise (as compared to the situation under the classical model, where a gene is removed from selection once it has become nonfunctionalized). Second, the partitioning of gene expression patterns by the DDC process may reduce the pleiotropic constraints operating on single-gene loci, thereby allowing natural selection to more closely tune the duplicate members of a pair to their specific subfunctions. Third, gene duplicates that have unresolved subfunctions at the time of a reproductive isolation event may provide a powerful mechanism for the development of reproductive incompatibility, *i.e.*, speciation. The degeneration of orthologues in different ways in two sister taxa effectively causes a divergence in genetic maps (Haldane 1933) and a consequent loss of some aspects of gene expression in hybrid progeny.

The purpose of this article is to evaluate the conditions under which duplicate-gene preservation by the DDC process is likely to be quantitatively significant. In our

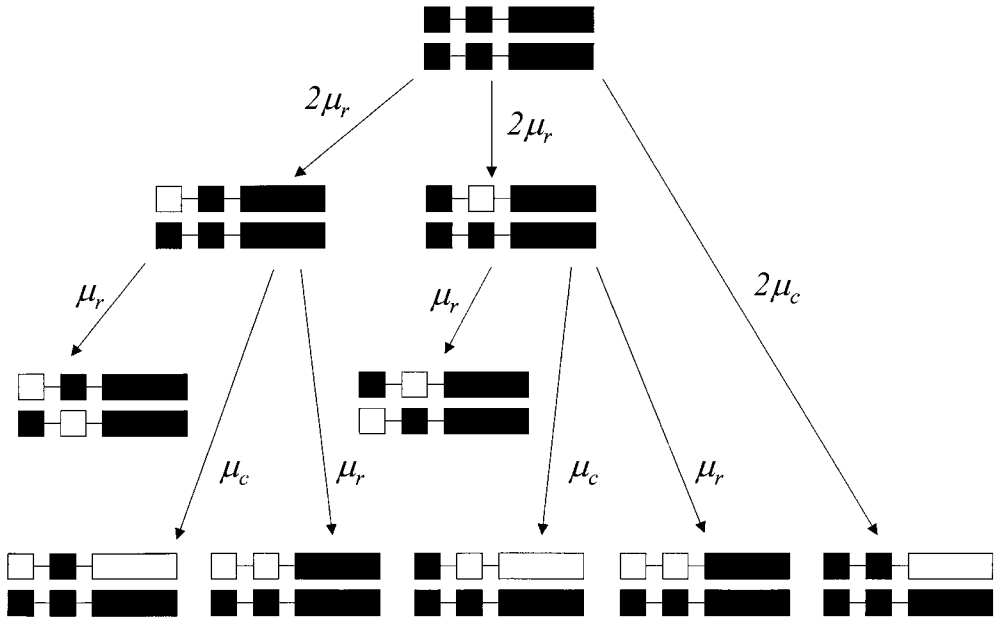

Figure 1.—The two possible fates of a pair of gene duplicates under the idealized subfunctionalization model. Here we assume a gene with two independently mutable subfunctions (depicted as regulatory regions by the two small boxes), which are spatially nonoverlapping with each other and with the coding region (depicted as a rectangle). μ_r is the rate at which a subfunction is knocked out by mutation, while μ_c is the rate at which complete loss-of-function mutations arise. Only the haploid states of the duplicate loci are shown. The states in the second row denote situations in which one member of the pair has lost one subfunction. The states in the third row denote gene preservation by subfunctionalization; here, the two copies have lost single, non-overlapping subfunctions, and therefore complement each other. The states in the fourth row denote gene loss by nonfunctionalization; this occurs when either the coding region is knocked out or both regulatory regions have been lost.

earlier work (Force *et al.* 1999), we presented an analytical approximation for the extreme situation in which individual fitness is only reduced when both alleles at both loci are null for a particular subfunction (the double-null recessive model). Our earlier study also assumed an effective population size that is small enough that the frequency of double-null homozygotes is negligible enough that mutant alleles drift to fixation in an effectively neutral manner, and we restricted our attention to the situation in which the two members of a pair of duplicate genes are freely segregating, as in the case of polyploid individuals that have become functionally diploidized. Here we evaluate how population size influences the probability of duplicate-gene preservation, examine the consequences of incomplete dominance and of partial loss-of-activity mutations, and investigate the evolutionary fates of tandem *vs.* freely segregating duplicates.

BACKGROUND

We initially focus on mutations that cause complete loss of some or all aspects of gene expression, returning later to consider the consequences of mutations with smaller effects. Here we have in mind events such as large insertions or deletions in regulatory regions or frameshift mutations in coding regions. An idealized scenario is laid out in Figure 1, which considers a gene consisting of a coding region and two independently mutable regulatory regions (one for each of two subfunctions). Each regulatory region incurs a knockout mutation for a specific subfunction with rate μ_r per gene per generation, whereas the coding region mutates to a completely nonfunctional allele at rate μ_c. The actual model pursued below is more general than the spatial arrangement implied in the figure, because μ_r is really just the rate at which single subfunctions are eliminated by mutation and μ_c is the rate of origin of complete loss-of-function alleles (mutations that simultaneously eliminate all subfunctions). The relative values of these two mutation rates may be only weakly correlated with the amount of DNA associated with coding and regulatory sequences. For example, long stretches of a protein can sometimes be removed with little effect on gene function, and the modular molecular architecture of some proteins results in tissue-specific effects of coding-region mutations (Henikoff *et al.* 1997). In addition, insertions well outside of transcription-factor binding sites may have substantial general or specific effects.

Because we are primarily interested in the extent to which duplicate-gene preservation can be understood in terms of degenerative mutation, we ignore rare beneficial mutations to new functions. Under this assumption, a pair of duplicate genes will then ultimately succumb to one of two possible fates: (1) nonfunctionalization occurs when one of the two loci experiences fixation of a null (complete loss-of-function) allele, either by the loss of the coding region or by the sequential loss of all of the subfunctions from one gene copy and (2) subfunctionalization occurs when the two duplicate loci become fixed for complementary loss-of-subfunction mutations, thereby resulting in their reciprocal preser-

vation. The probabilities of these alternative outcomes, which sum to one, are denoted P_n and P_s, where n stands for nonfunctionalization and s for subfunctionalization.

The development of a general analytical model to predict P_s presents formidable technical challenges. For example, even when there are only two subfunctions for a gene, there are four possible classes of alleles: fully functional alleles, mutant alleles retaining only the first subfunction or only the second subfunction, and null alleles (Figure 1). There are then 10 genotypes at each locus and 100 two-locus genotypes. With three subfunctions in the original gene, there are eight classes of alleles: fully functional alleles with all three subfunctions intact, three types with only two subfunctions intact, three types with a single subfunction intact, and null alleles. More generally, with n subfunctions, there are 2^n classes of alleles and $[2^{n-1}(2^n + 1)]^2$ two-locus genotypes.

Because of the two-locus, multiallelic nature of the DDC process, most of our analyses have relied on computer simulations. In our initial studies, all such work was performed with individual-based simulations, wherein each offspring genotype was produced by randomly drawing a pair of parents, obtaining the gametes by random segregation and recombination, and imposing mutations stochastically according to the Poisson distribution. Because these types of simulations can be very time-consuming for large population sizes, for the case of two subfunctions per gene, we ultimately settled on an alternative approach that simply kept track of genotype frequencies. With this approach, an effectively infinite gamete pool was assumed, so that conditional on the parental genotype frequencies, recombination and mutation could be treated as deterministic processes in the production of the gamete pool for the next generation. The gamete frequencies were then used to derive the expected genotype frequencies after random mating and selection. Using these expectations, the actual genotype frequencies were obtained by sequential binomial sampling. This second approach gave results that were indistinguishable from those obtained with the individual-based model.

All of the results reported below assume a constant population size (N) and a monoecious mating system, and in most cases (all cases in which $N \leq 10^4$), 1000 simulations were performed for each set of parameters. Throughout, we assumed a loss-of-function mutation rate of $\mu_c = 10^{-5}$ per allele per generation.

THE DOUBLE-NULL RECESSIVE MODEL

The classical model: The classical model can be viewed as a special case of the DDC model in which there is only a single function per gene, and its analysis provides a useful basis for comparison with genes with multiple subfunctions. Most of the work on the classical model of gene duplication has considered the situation in which mutant alleles are completely recessive. Under the double-null recessive model, all individuals with at least one functional gene are fully viable, whereas homozygotes for null alleles at both loci have zero fitness. In this case, there is nothing to prevent the ultimate loss of the active allele at one locus in a finite population, so the only issue is the length of time for this to occur. Although a number of studies have attempted to answer this question (Bailey *et al.* 1978; Kimura and King 1979; Takahata and Maruyama 1979; Li 1980), the results are somewhat inconsistent, perhaps because of the small number of replications in some of the computer simulations. However, Watterson (1983) subsequently used diffusion theory to derive an estimator for the mean time to nonfunctionalization for a member of a pair of unlinked gene duplicates. His formula can be expressed as

$$\bar{t}_n = N\left[\log(2N) + 0.57721 + \left(1 - \frac{\theta}{2}\right)\sum_{i=1}^{\infty}\frac{1}{i[(\theta/2) + i - 1]}\right], \quad (1)$$

where $\theta = 4N\mu_c$ and μ_c is the rate of mutation to null alleles.

Because the behavior of Watterson's formula has never been thoroughly examined, we compared its performance with results generated by computer simulation. As can be seen in Figure 2, Equation 1 yields predictions that are in remarkably good agreement with simulated data for freely recombining loci. For $\mu_c N < 0.1$, the mean time to nonfunctionalization is slightly greater than $1/(2\mu_c)$ generations, but for $\mu_c N > 1$, $\bar{t}_n$ is prolonged to roughly $10N$ generations as selection becomes more efficient. The mean time to nonfunctionalization for completely linked gene duplicates is essentially the same as that for unlinked loci when $\mu_c N < 0.1$, but is reduced when $\mu_c N$ is larger (Figure 2). Thus, under the classical double-null recessive model, for populations with effective sizes less than $1/\mu_c$, we expect silencing of one member of a duplicate pair to occur in less than a million generations or so, whereas extremely large populations may harbor active pairs of gene duplicates for tens of millions of generations.

Two subfunctions: As noted above, when a gene duplicate has independently mutable subfunctions, the possibility exists that the two members of the pair may experience fixation of complementary loss-of-subfunction mutations, leading to gene preservation rather than gene loss. For sufficiently small population sizes, we may expect fixation events to occur at the two loci in a nonoverlapping manner and double-null homozygotes to be rare enough that selection is essentially inoperable. The evolutionary fate of the duplicate pair can then be approximated by a neutral model, which simply tallies the alternative series of mutational events that can occur at the two loci (Force *et al.* 1999). We briefly review our derivation of the probability of subfunctionalization, P_s, under the assumptions of effective neutrality for the case of a gene with two subfunctions.

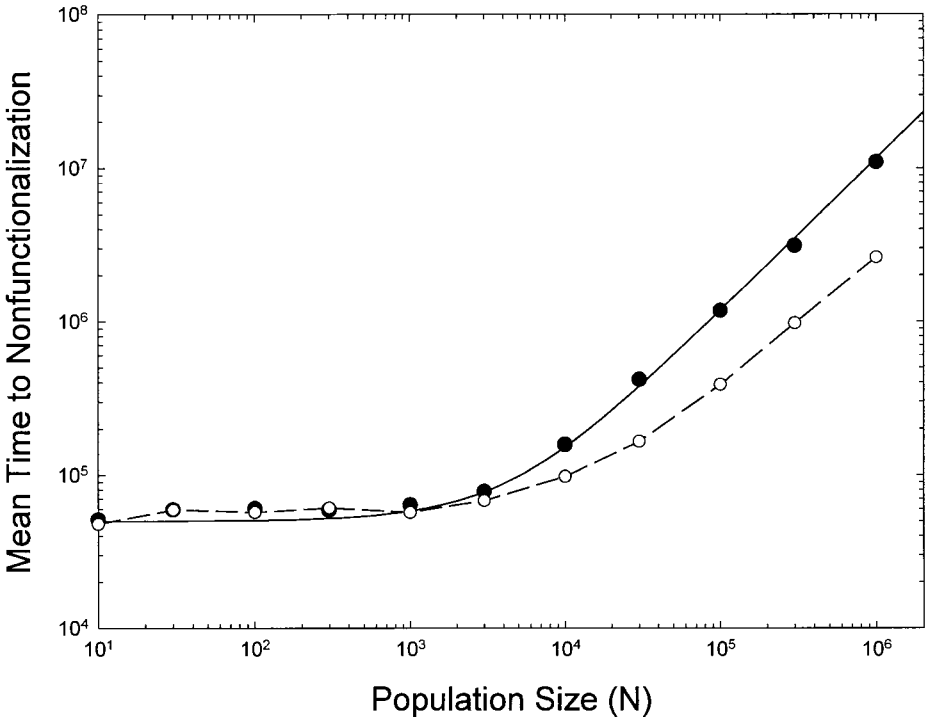

Figure 2.—The mean time to gene silencing under the classical model of gene duplication. The solid (free recombination) and open (complete linkage) circles are each average results from ~1000 simulations, with a mutation rate equal to 10^{-5} per gene per generation. The solid line is generated by use of Equation 1, whereas the dashed line simply connects the observed data for completely linked duplicates.

Consider the situation in which each subfunction is subject to mutational loss at the rate μ_r, and let μ_c be the rate at which complete nonfunctionalizing mutations arise. The mutation rate for a completely functional gene is then $\mu_c + 2\mu_r$ per gene copy. Under the assumption of effective neutrality, the rate of fixation of a mutation at a locus is equal to the genic mutation rate (Kimura 1983), so the probability that the first fixation event does not lead to the production of a pseudogene is equal to the total rate of subfunctionalizing mutations divided by the total mutation rate for the gene, *i.e.*, $2\mu_r/(\mu_c + 2\mu_r)$. Given the elimination of one of the subfunctions from the first gene copy, the second copy must maintain this subfunction, as complete loss of an essential expression domain is assumed to be lethal. Thus, the rate of origin of mutations in the second copy that can subsequently become fixed is now reduced to μ_r, whereas additional fixable null mutations in the partially degraded first copy can occur both in the remaining regulatory subfunction and in the coding region. Therefore, the total rate (summed over both copies) for the second fixation event is $(\mu_c + 2\mu_r)$. The probability of gene preservation by subfunctionalization is equal to the probability that the coding regions have survived the first hit, $2\mu_r/(\mu_c + 2\mu_r)$, multiplied by the probability that the second mutation occurs in a complementary subfunction in the second copy, $\mu_r/(\mu_c + 2\mu_r)$,

$$P_s = 2\left(\frac{\mu_r}{\mu_c + 2\mu_r}\right)^2. \qquad (2)$$

The results from computer simulations, for both linked and unlinked loci, demonstrate that the probability of subfunctionalization (P_s) is essentially independent of population size and adequately approximated by Equation 2 provided $N(\mu_c + \mu_r) < 0.1$ (Figure 3). At larger $N(\mu_c + \mu_r)$, P_s for linked duplicates can greatly exceed that for unlinked loci. However, the range of values of $N(\mu_c + \mu_r)$ for which this is true is fairly restrictive, as the probability of subfunctionalization drops off fairly rapidly beyond the point at which $N(\mu_c + \mu_r) \simeq 0.1$. Even for completely linked genes, there is essentially no chance of preservation of gene duplicates under the double-null recessive model when $N(\mu_c + \mu_r) > 10$.

Why does the probability of preservation of duplicate genes by subfunctionalization decline at high $N(\mu_c + \mu_r)$? Even where it can be reasonably assumed that selection is negligible in determining the fates of gene duplicates, because the average time to fixation of a neutral gene is ~$4N$ generations (Kimura and Ohta 1969), when $N(\mu_c + \mu_r)$ is on the order of 0.1 or greater, there is an appreciable probability that all descendants of a mutant allele that is destined to fixation will acquire secondary mutations (either directly or through inheritance) during their sojourn through the population. Thus, when $N(\mu_c + \mu_r)$ is large, an initially subfunctionalized allele (and all of its descendants in the gene genealogy) may become silenced by secondary mutations during the fixation process, thereby increasing the probability that nonfunctionalization will be the ultimate fate of a duplicate pair of genes. We refer to this consequence of secondary mutation as mutational conversion.

Obtaining an approximation for P_s that incorporates secondary mutations is fairly straightforward for the case in which a gene has only two subfunctions, because

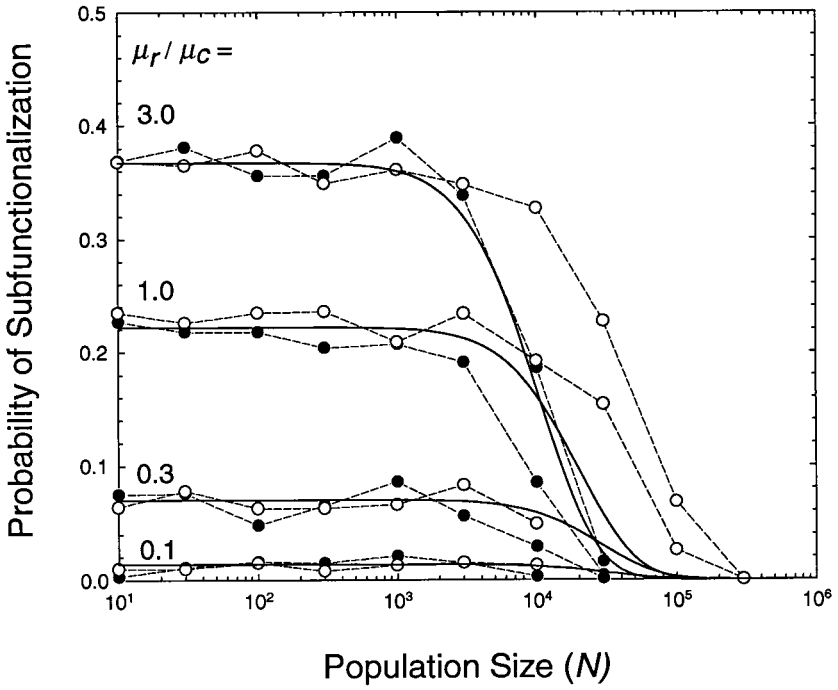

Figure 3.—The probability of subfunctionalization for a pair of gene duplicates (with two subfunctions) as influenced by population size. Results are given for four ratios of subfunctionalizing to nonfunctionalizing mutation rates (μ_r/μ_c). The solid and open points represent, respectively, computer-simulation results for the cases of free recombination and complete linkage. The solid lines are the analytical approximations provided by Equation 3. A coding-region mutation rate of $\mu_c = 10^{-5}$ per generation is assumed throughout.

in this instance all secondary mutations lead to gene silencing. Letting c be the probability that all descendants of a subfunctionalized allele destined to fixation are rendered nonfunctional by the time the lineage fixes, then from the arguments given above, the probability that the first fixation event involves the loss of one subfunction from one of the copies is reduced to $2\mu_r(1 - c)/(\mu_c + 2\mu_r)$. Given that this occurs, there are two possible outcomes in the next stage—the coding region or the alternative subfunction of the partially degraded first copy may be knocked out (at rate $\mu_c + \mu_r$), thereby leading to nonfunctionalization, or the second copy may lose the subfunction that remains intact in the first copy [at rate $\mu_r(1 - c)$], thereby resulting in subfunctionalization. Thus, a more general expression for the probability of subfunctionalization for genes with two subfunctions is

$$P_s = \left(\frac{2\mu_r(1 - c)}{\mu_c + 2\mu_r}\right)\left(\frac{\mu_r(1 - c)}{\mu_c + (2 - c)\mu_r}\right), \qquad (3)$$

which reduces to Equation 2 when $c = 0$.

In the appendix, we derive an approximate expression for the probability of mutational conversion (c) for genes with two subfunctions using a gene genealogical approach and known properties of the coalescent for neutral genes. A useful property of the resulting theory is that c depends simply on the product $N(\mu_c + \mu_r)$ (Figure 4). For $N(\mu_c + \mu_r) < 0.01$, the probability of mutational conversion is essentially zero, whereas for $N(\mu_c + \mu_r) > 5$, virtually all subfunctionalized alleles that are destined for fixation are expected to be converted to nonfunctional alleles in transit. The decline in P_s observed with increasing N is in rough accord with the predictions of Equation 3 for unlinked genes (Figure 3). Equation 3 tends to overestimate P_s in the region of $0.1 > N(\mu_c + \mu_r) > 1$, perhaps because selection plays a small role in this region. However, the neutral theory does provide a fairly good indication of the population size beyond which the likelihood of gene preservation by subfunctionalization is vanishingly small.

As in the case of the classical model, the mean time to resolution (either by nonfunctionalization or subfunctionalization) under the DDC model is generally on the order of $1/(2\mu_c)$ when $N\mu_c < 0.1$, and is more on the order of $10N$ generations for larger $N\mu_c$ (Figure 5). Although the mean time to resolution depends somewhat on the ratio μ_r/μ_c when N is small, this dependence is not strong, and for unlinked loci Equation 1 provides a good qualitative approximation for the full range of N. Given the previous results with the classical model, the agreement with the theoretical expectation is expected to be very good when N is large, because in this case the outcome is identical to that under the classical model—all pairs of gene duplicates are resolved by nonfunctionalization of one member of the pair.

Additional subfunctions: Under the DDC model, gene preservation is expected to increase with the number of independently mutable subfunctions, because this increases the number of pathways by which complementation can occur. For example, with three subfunctions, gene preservation can occur in two steps by three different pathways—with any one of the three subfunctions first being eliminated from one copy, followed by

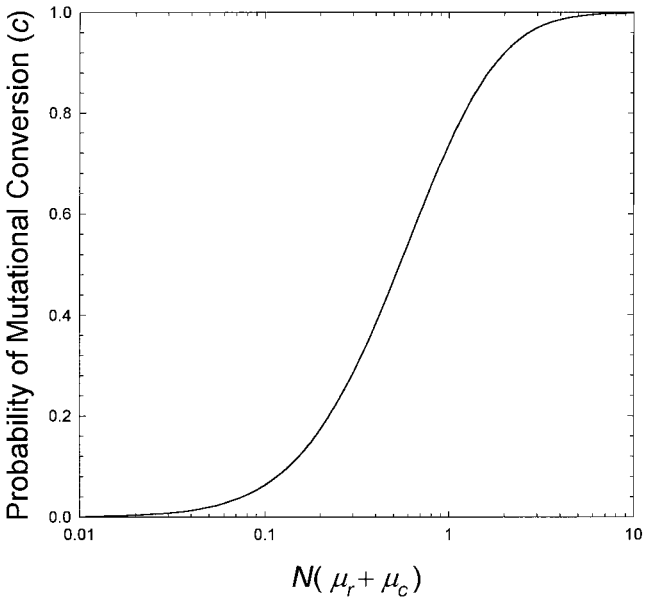

Figure 4.—The probability of mutational conversion from an allele that has lost one of two subfunctions to an allele that is completely nonfunctionalized. N is the population size, μ_c is the rate of mutation to loss-of-function alleles, and μ_r is the mutation rate for single subfunction loss. The relationship given by the solid line was obtained using the solution provided in the appendix.

loss of a different subfunction from the second copy; or in three steps by three different pathways—with any two of the three subfunctions first being eliminated from one copy, followed by loss of the third subfunction from the second copy.

Under the assumption of effective neutrality, we have derived the generalization of Equation 2 for genes with an arbitrary number (z) of subfunctions (Force *et al.* 1999). The probability that gene preservation occurs by a pathway involving i steps, *i.e.*, with $(i-1)$ consecutive fixations of subfunctionalizing mutations on one copy followed by one on the other, is given by

$$P_{s,i} = \left(\frac{z\mu_r}{\mu_c + z\mu_r}\right)\prod_{j=0}^{i-2}\left(\frac{(z-j-1)\mu_r}{\mu_c + 2(z-j-1)\mu_r}\right), \quad (4)$$

and the total probability of gene preservation by subfunctionalization is then obtained by summing this quantity over $i = 2$ to z,

$$P_s = \sum_{i=2}^{z} P_{s,i}. \quad (5)$$

As can be seen in Figure 6, this approximation works very well when $N(\mu_c + z\mu_r) < 0.01$ for all z. With even a moderate number of independently mutable subfunctions, P_s can become quite high when populations are small to moderate in size. For example, with $\mu_r/\mu_c = 1$, P_s at low N asymptotically approaches 0.22 with two subfunctions, 0.40 with three subfunctions, and 0.61 with five subfunctions (Figure 6). For population sizes in the range $0.1 > N(\mu_c + \mu_r) > 1$, P_s can actually exceed the neutral expectation when z is large, presumably because the efficiency of selection against nonfunctional alleles is increased when large numbers of subfunctionalized alleles are segregating. Eventually, however, regardless of the number of subfunctions, P_s declines to zero as N becomes very large and mutational conversion prevents the fixation of subfunctionalized alleles.

Some insight into the upper limit to the probability of subfunctionalization can be obtained by treating the ratio of the total subfunctionalization mutation rate to the nonfunctionalization rate, $r = z\mu_r/\mu_c$, as a constant,

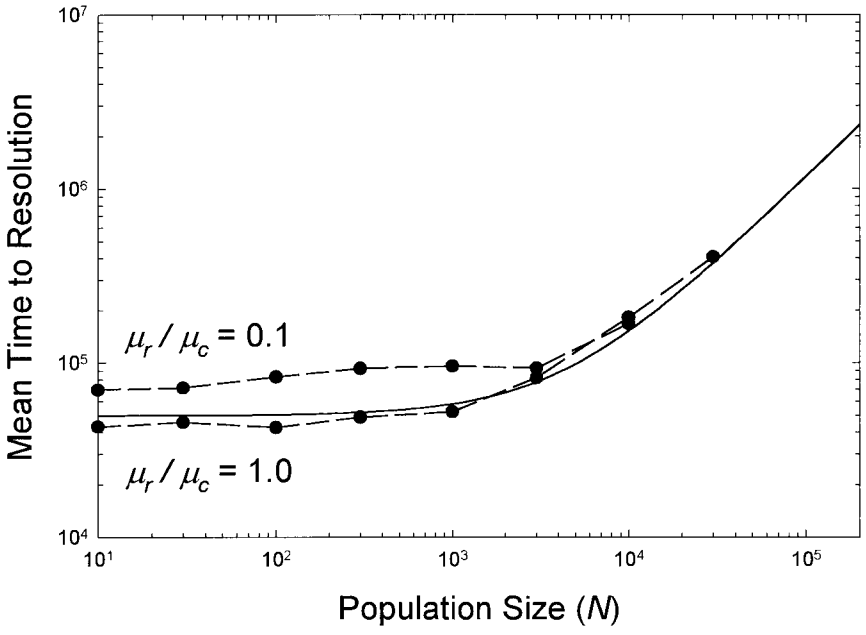

Figure 5.—The mean time to duplicate-gene resolution (by nonfunctionalization or subfunctionalization) for a pair of unlinked duplicates with two subfunctions. Simulation results are given for two ratios of μ_r to μ_c, with $\mu_c = 10^{-5}$ in all cases. The solid line gives the predictions from Equation 1 with $\mu_c = 10^{-5}$.

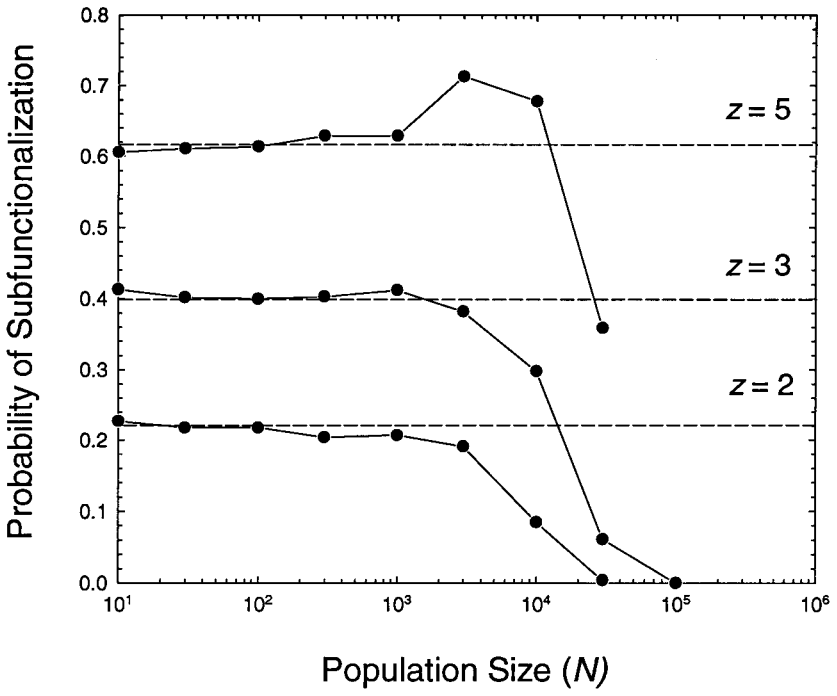

Figure 6.—The probability of subfunctionalization for a pair of gene duplicates with a ratio of subfunctionalizing to nonfunctionalizing mutation rates of $\mu_r/\mu_c = 1$ as a function of population size. The two genes are assumed to be unlinked, and $\mu_c = 10^{-5}$. Most data points are the average results of 1000 computer-simulation runs. Results are given for genes with two, three, and five independently mutable subfunctions. The dashed lines are the small-population-size approximations provided by Equation 5.

and evaluating the limit of Equation 5 as $z \rightarrow \infty$, *i.e.*, as the number of targets for subfunctionalization becomes effectively infinite,

$$P_s = \left(\frac{z\mu_r}{z\mu_r + \mu_c}\right)^2. \qquad (6)$$

Thus, if $r = 0.5$, the probability of subfunctionalization under the double-null recessive model is no greater than 0.11, whereas the upper limit to P_s is 0.25 with $r = 1$ and 0.44 with $r = 2$. For $r \ll 1$, $P_s \simeq r^2$.

PARTIAL DOMINANCE

In the previous section, we focused on the situation in which null alleles are completely recessive with respect to fitness. We now examine the situation in which two (rather than the previous one) active alleles for each subfunction are required, with individuals having one or zero active alleles for any subfunction being inviable. Such a condition is often referred to as haplo-insufficiency, although we allow the two alleles to be present at either locus. Under this model, heterozygotes at a particular locus are selectively eliminated whenever they appear on a background of a null homozygote at the alternative locus, so one would expect the time to resolution of the fates of gene duplicates to be extended. This, in fact, is observed, although the effect is not large (data not shown). Of greater interest is the relative insensitivity of the probability of subfunctionalization to dosage requirements. Except for the narrow range of population sizes in which P_s rapidly declines to zero, P_s is essentially the same under both the double-null recessive and haplo-insufficiency models (Figure 7).

PARTIAL LOSS-OF-FUNCTION OR PARTIAL LOSS-OF-SUBFUNCTION MUTATIONS

In all of the preceding analyses, we assumed that mutations completely eliminate all activity of a single subfunction or of both subfunctions. We now show that this extreme assumption leads to minimum predicted levels of duplicate-gene preservation, confining our attention to population sizes that are small enough to fulfill the assumptions of effective neutrality. When mutations cause only a partial reduction in function, duplicate genes can be preserved by quantitative complementation (Force *et al.* 1999; Stoltzfus 1999), whereby the two copies must be maintained in the genome once the summed activity for a particular subfunction in both copies has been reduced to the original level in the single ancestral gene. We first consider how duplicate genes can be permanently preserved even in the absence of independent subfunctions, *i.e.*, if each copy is partially degraded such that the joint expression of both copies is necessary to fulfill functional requirements.

Assuming an additive model of gene action between loci and letting s be the number of degradational steps between full and no function, the probability of duplicate-gene preservation is equivalent to the probability that both copies will experience partial loss-of-function prior to the occurrence of complete loss-of-function from either copy,

$$P_s = \left(\frac{\mu_p}{\mu_p + \mu_c}\right)\sum_{j=1}^{s-1}\left(\frac{\mu_p}{2\mu_p + \mu_c}\right)^j, \qquad (7a)$$

where μ_p is the rate of occurrence of partial loss-of-function mutations, and μ_c is the rate of occurrence of

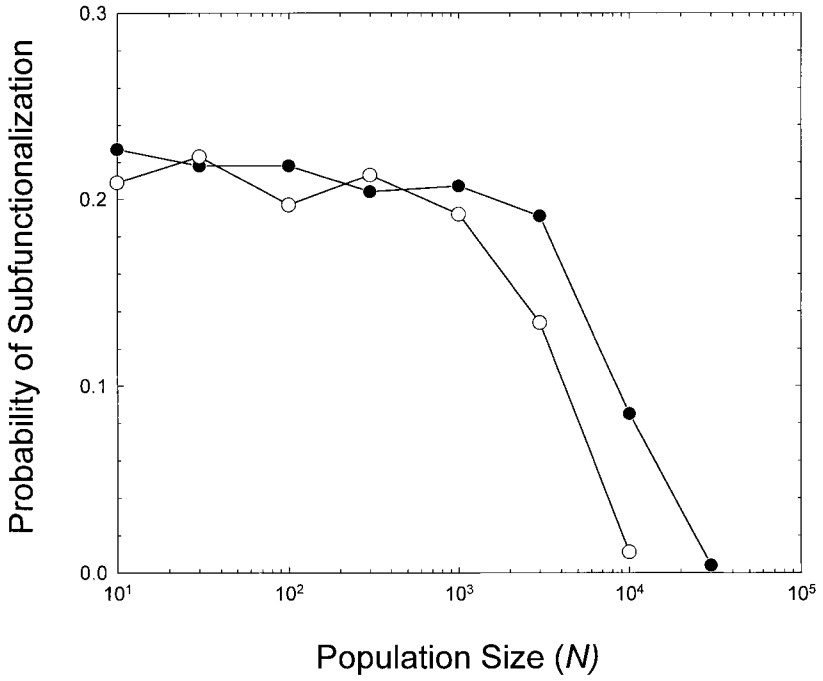

Figure 7.—The probability of subfunctionalization for a pair of unlinked gene duplicates with a ratio of subfunctionalizing to nonfunctionalizing mutation rates of $\mu_r/\mu_c = 1$ as a function of population size; $\mu_c = 10^{-5}$. The solid circles denote results for the double-null recessive model, whereas the open circles denote the results for the haplo-insufficiency model.

complete loss-of-function mutations. The upper limit to P_s approached as $s \to \infty$ is

$$P_s = \left(\frac{\mu_p}{\mu_p + \mu_c}\right)^2, \qquad (7b)$$

and this is closely approximated when $s > 5$ (Figure 8). Provided $\mu_p/\mu_c > 1$, which seems likely, the probability of duplicate-gene preservation by partial loss-of-function is substantial, even when s is as small as 2 (Figure 8). For $\mu_p/\mu_c > 10$, P_s is closely approximated by $1 - (1/2)^{s-1}$.

We next consider the case in which mutations to two independently mutable subfunctions partially reduce activity, whereas those to the coding region completely eliminate function. Under the model considered above in Equations 2 and 3, there was only one path to gene preservation by subfunctionalization (the complete loss of one subfunction from one copy, followed by the complete loss of the second subfunction from the second copy). However, even with mutations that reduce subfunction by 50%, there are six different paths to duplicate-gene preservation (Figure 9). The probability of each path is obtained by multiplying the chain of relevant transition probabilities, and the total probability of duplicate-gene preservation for this $s = 2$ case is simply the sum of the probabilities of the six paths,

$$P_s = r_0 r_1[1 + 0.5(r_1 + r_2)(1 + r_2)], \qquad (8a)$$

where $r_0 = 2\mu_r/(2\mu_r + \mu_c)$, $r_1 = 2\mu_r/(4\mu_r + \mu_c)$, and $r_2 = \mu_r/(2\mu_r + \mu_c)$. For this model with $s = 3$,

$$P_s = r_0 r_1\left(1 + r_1 + \frac{r_1}{4}(3r_1 + r_2) + \frac{r_1}{8}(3r_1^2 + 3r_1 r_2 + 2r_2^2)(1 + r_2)\right), \qquad (8b)$$

and the asymptotic limit as $s \to \infty$ is

$$P_s = \left(\frac{2\mu_r}{2\mu_r + \mu_c}\right)^2. \qquad (8c)$$

For this case of two subfunctions, the probability of gene preservation can be increased as much as twofold when mutations have partial degenerative effects, and almost all of the increase is realized when two, rather than one, mutations are required for the complete loss of a subfunction (Figure 10). The expectations for $s = 3$ are not very different from those for $s = \infty$.

A general approximation for an arbitrary number of subfunctions (z) and an arbitrary number of steps to complete silencing (s) is given by

$$P_s = \left(\frac{1 - (1/2)^{zs-1}}{1 - (1/2)^{z-1}}\right)P_s', \qquad (9)$$

where P_s' is given by Equations 2 and 5. The denominator of the fraction on the left is the asymptotic value of P_s that is approached with high μ_r/μ_c when subfunction-specific mutations lead to complete loss ($s = 1$), whereas the quantity in the numerator is the asymptotic value for arbitrary s. Thus, contrary to the situation with $z = 2$, where P_s can be increased by as much as 50% with partial loss-of-subfunction, with $z = 3$ the maximum increase is 33%, and with $z = 5$ it is only 7% (Figure

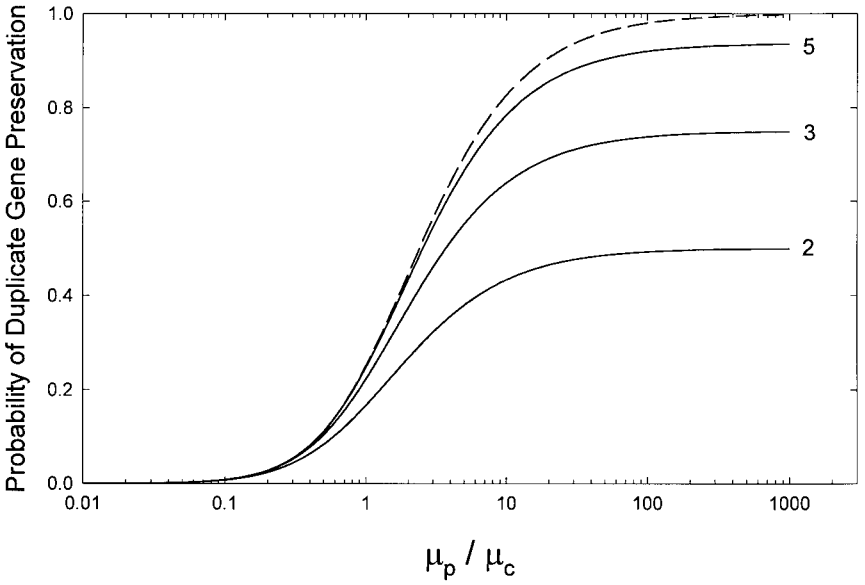

Figure 8.—The probability of duplicate-gene preservation for the case in which there are no independent subfunctions, as a function of μ_p/μ_c, the ratio of partial to complete loss-of-function mutations, and of s, the number of partial loss-of-function mutations required to completely silence a gene. Results are given for $s = 2$, 3, and 5 (solid lines). The dashed line is the asymptotic limit as $s \to \infty$.

10). An intuitive explanation for this behavior is that when the number of subfunctions is even moderately high, almost all cases of complete gene silencing are a consequence of coding-region nulls (rather than of multiple eliminations of subfunctions), and further increasing the number of paths to subfunctionalization by increasing s does not appreciably change the situation.

The examples presented above cover only some limiting situations, with more complex scenarios leading to even higher rates of duplicate-gene preservation. For example, as noted above (Figure 8), if an appreciable fraction of mutations arising in the coding region (or more generally, mutations that jointly influence all subfunctions) lead to partial, rather than full, loss of expression, then the probability of gene preservation at low μ_r/μ_c will not be zero. Instead, it will asymptote at values close to those illustrated in Figure 8 for various ratios of partial to complete loss-of-function mutation rates in the coding region. These results indicate that when the average effects of mutations influencing both complete

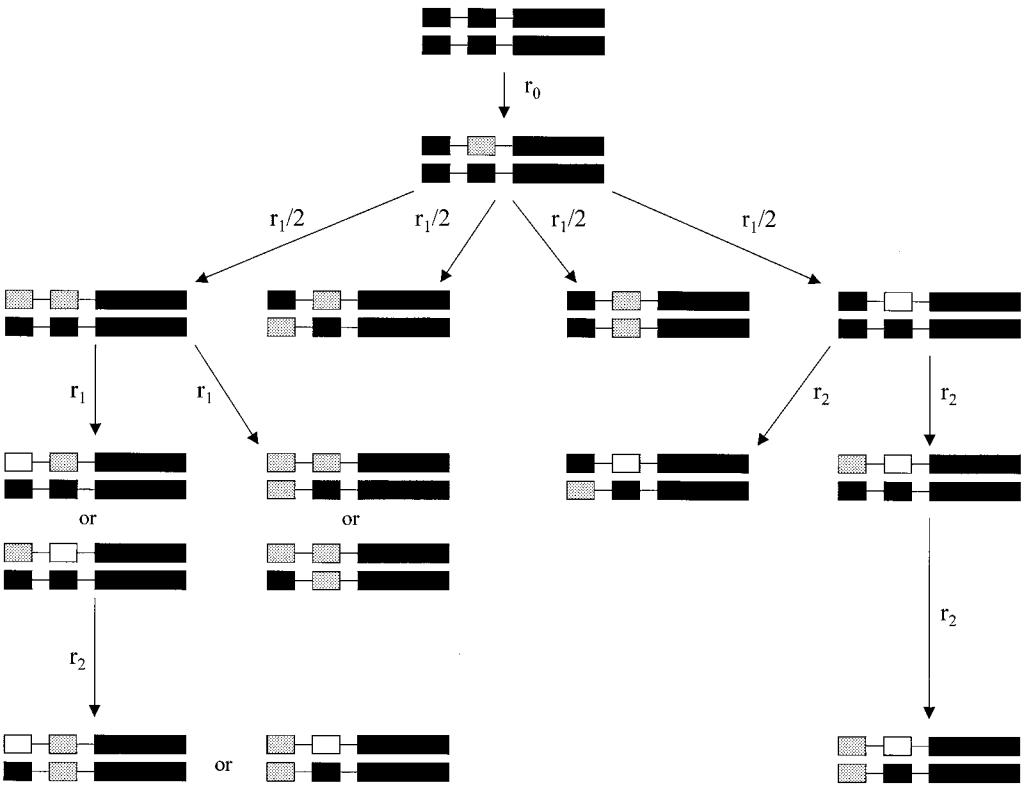

Figure 9.—The six paths to gene preservation by degenerative mutation when there are two independently mutable subfunctions and mutations to such subfunctions cause 50% loss-of-function, whereas coding-region mutations cause complete loss-of-function. Black denotes a region of the gene that is mutation free; gray denotes a subfunction that has been hit with one mutation, and an open box denotes a twice-hit region, for which there is no remaining activity. The coefficients denote transition probabilities as defined in the text. The several paths to nonfunctionalization of one copy are not shown.

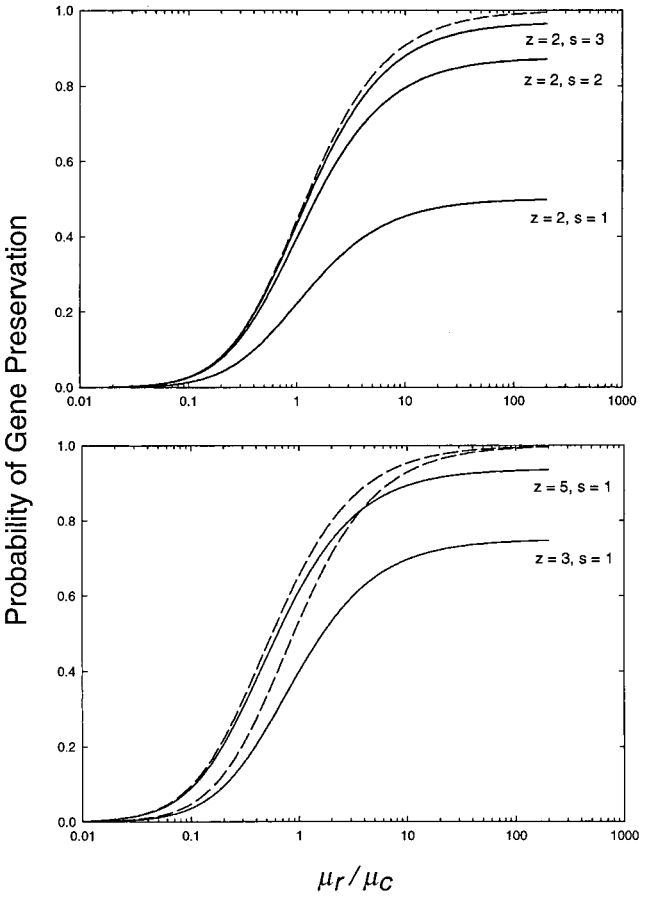

Figure 10.—The expected probability of gene preservation for the case in which $N(2\mu_r + \mu_c) < 0.1$, for various ratios of the mutation rate to subfunctionalizing *vs.* nonfunctionalizing mutations. (Top) Four cases involving two subfunctions are illustrated, the solid lines denoting P_s when one, two, or three mutations are required for complete loss of a subfunction, and the dashed line denoting the asymptotic limit for mutations with small effects. (Bottom) Cases in which there are three or five subfunctions are illustrated, the solid lines denoting the situation when single mutations are sufficient for the complete loss of a subfunction, and the dashed lines denoting the asymptotic limits for mutations with small effects (the upper and lower curves representing the cases for $z = 5$ and 3, respectively).

function and individual subfunctions are high, the probability of duplicate-gene preservation by degenerative mutations can be very substantial.

DISCUSSION

Under the classical model of gene duplication, nonfunctionalization of one member of the pair by degenerative mutation has generally been viewed as inevitable unless the fixation of a silencing mutation is preceded by a mutation to a novel beneficial function. However, there now appear to be several plausible mechanisms for the preservation of duplicate genes (Clark 1994; Nowak *et al.* 1997; Force *et al.* 1999; Stoltzfus 1999; Wagner 1999). The DDC model postulates that degenerative mutations result in the preservation of gene duplicates through the production of loci with complementing sets of subfunctions. Because degenerative mutations are much more frequent than beneficial mutations and because many genes have complex regulatory regions driving tissue-specific patterns of expression, it follows that subfunctionalization may be a much more common mechanism of duplicate-gene preservation than neofunctionalization. If this idea is correct, then the mechanism that results in the preservation of gene duplicates in the genomes of complex organisms is distinct from the subsequent mechanisms that result in the origin of new gene functions. There is, however, nothing inherent in the DDC model that denies the significance of gene duplication in the origin of evolutionary novelty. Indeed, the subfunctionalization process may facilitate such evolution by preserving gene duplicates and maintaining their exposure to natural selection and/or by removing pleiotropic constraints.

The results presented above help clarify the conditions under which duplicate-gene preservation by subfunctionalization is likely to be quantitatively significant. First, subfunctionalization is most likely to occur when the effective population size and the coding null mutation rate is low enough that a new loss-of-function mutation arises in the population every five generations or less, *i.e.*, roughly speaking, when $N\mu_c < 0.1$. Moreover, provided these conditions are met, the probability of gene preservation by subfunctionalization can be fairly accurately predicted by use of a model that largely ignores the details of selection. The apparent reason for this behavior is that when the number of mutations arising per generation is small, the very low frequency of individuals with degenerative mutations at both loci results in dynamics of gene frequency change that are largely a consequence of random genetic drift and the relative incidence of different mutational types, *i.e.*, the evolutionary fates of mutations at each locus are essentially independent of the alleles segregating at the other locus. It is also for this reason that the results for the double-null recessive and haplo-insufficiency models are essentially identical for $N\mu_c < 0.1$. For vertebrates, the condition $N\mu_c < 0.1$ is probably not uncommon when one considers that μ_c is on the order of 10^{-6}–10^{-5}, that the long-term effective size of a population is on the order of the minimum annual effective size, and that the effective size of a population is often on the order of one-tenth to one-third of the actual number of breeding adults.

On the other hand, the subfunctionalization process appears to be an unlikely mechanism of duplicate-gene preservation when $N(\mu_r + \mu_c) > 10$. With very large population sizes, selection begins to play a more significant role, so the probability of fixation of *any* mutant allele is expected to be diminished and the time to fixation to be magnified. Based on the observations for

the classical model, under which fixation of a null allele at one locus must ultimately occur, the average time until one locus becomes silenced is on the order of $10N$ generations when N is large (Figure 2), and even for neutral alleles the time to fixation is approximately $4N$ generations. Thus, when $N \gg 1/(\mu_r + \mu_c)$, any mutant allele that is destined to fixation is likely to acquire secondary mutations in transit, and once the population size becomes very large, mutational conversion of subfunctionalized to nonfunctionalized alleles eliminates any possibility of subfunctionalization being the ultimate fate of duplicate genes. Roughly speaking, $N(\mu_r + \mu_c) > 10$ implies an effective population size on the order of 10^6 to 10^7 or greater, so population-size conditions that completely thwart the subfunctionalization process are not necessarily common. In contrast to the situation with subfunctionalization, neofunctionalization or positive selection for redundancy appears to be ineffective at preserving gene duplicates at small population sizes, only becoming plausible explanations at effective population sizes on the order of 10^6 or greater (Clark 1994; Walsh 1995; Nowak *et al.* 1997; Wagner 1999).

Second, our results suggest that provided $N(\mu_r + \mu_c) < 0.1$, the degree of linkage between two gene duplicates plays a negligible role in their ultimate fate, *i.e.*, the probability of preservation for tandem duplicates is essentially the same as it is for duplicates carried on different chromosomes. Such behavior is expected under these conditions because mutant alleles are rare enough that they have essentially no influence on the dynamics of gene frequency at the opposite locus. On the other hand, when $0.1 < N(\mu_r + \mu_c) < 10$, a completely linked pair of duplicates has a higher probability of preservation by the DDC process than an unlinked pair. This elevated probability of subfunctionalization of linked *vs.* unlinked duplicate genes at intermediate population sizes is very likely a consequence of the Hill-Robertson effect, whereby linked deleterious genes interfere with each other's selective elimination (Hill and Robertson 1966; Birky and Walsh 1988). Such selective interference is consistent with the reduced time to silencing of linked duplicates under the classical model (Figure 2), and under the subfunctionalization model a reduced time to fixation translates further into a reduced probability of mutational conversion to complete nulls. Finally, once $N(\mu_r + \mu_c) > 10$, mutational conversion plays such a dominant role that duplicate-gene preservation by the DDC process is negligible regardless of the degree of linkage. Thus, the range of effective population sizes over which P_s is expected to differ between tandem duplicates and unlinked duplicates appears to be relatively small. It should be noted, however, that we have only examined the effects of intergenic, not intragenic, recombination.

Third, genes with as few as three to five independently mutable subfunctions are expected to have greatly elevated probabilities of gene preservation by the DDC process relative to the case in which there are only two subfunctions. Part of the reason for this behavior is simply that a greater number of independent regulatory-region targets lead to a greater number of paths by which a gene pair can be subfunctionalized. As shown by the limits of subfunctionalization (Equations 6, 7b, and 8c), the relative rate of subfunctionalizing mutations is far more important than the actual number of subfunctions. However, with a larger number of independently mutable subfunctions, the probability of mutational conversion of a subfunctionalized allele to a complete null is also reduced. Thus, as we have pointed out earlier (Force *et al.* 1999), a fairly robust prediction of the DDC model is that the probability of duplicate-gene preservation will be higher in genes with greater regulatory-region complexity, and as a corollary, will be reduced over subsequent rounds of duplications (as the different members of the pair progressively lose the subfunctions that would otherwise foster preservation).

Fourth, whereas most of our results were obtained under the assumption that all mutations completely eliminate a function or subfunction, it is clear that mutations with smaller effects (even those that still cause as much as 50% loss of activity) can lead to a substantial increase in the probability of duplicate-gene preservation. As also pointed out by Stoltzfus (1999), this is even true for genes with a single function. When coding-region mutations [or overlapping or embedded regulatory regions; Force *et al.* (1999)] lead to only partial reduction in gene function, the increase in P_s can be especially dramatic. These outcomes arise when mutations have small degenerative effects because the probability that one member of a pair will be degraded to a completely nonfunctional form before the other has also been compromised becomes diminishingly small. Because it is likely that a substantial fraction of degenerative mutations (perhaps the majority) do not lead to complete loss-of-function (subfunction), this result further substantiates the argument that degenerative mutation may be the predominant mechanism that drives the accumulation of gene duplicates in developmentally complex organisms.

Fifth, we note that as the number of independently mutable subfunctions (z) and/or the number of degradational steps (s) becomes even moderately large (four or greater), the limit to the probability of subfunctionalization closely approaches the same form. As can be seen by comparing Equations 6, 7b, and 8c, the upper limit to P_s is in all cases equal to the square of the fraction of mutations that have partial effects. Thus, a key determinant of the role of degenerative mutations in the preservation of duplicate genes is the frequency of mutations with partial effects relative to that of complete nulls. Insight into this parameter should be achievable through mutation screens of alleles with known expres-

sion patterns and the analysis of interactions between them.

Finally, we note that we have couched all of the theory in this article in terms of loss-of-function/subfunction mutations. However, some evidence suggests that gain-of-function mutations may be quite common, perhaps as common as loss-of-function mutations (Clark *et al.* 1995). A key feature of the DDC model is how mutations in duplicate genes are perceived at the level of natural selection, so the extent to which our results would have to be modified in the face of gain-of-function mutations depends on their influence on fitness. If such mutations are dominant and have a negative influence on fitness, then they may be effectively purged from the population and contribute little to the long-term evolutionary dynamics of gene duplicates. On the other hand, gain-of-function mutations may prolong the life of gene duplicates by resurrecting previously impaired copies. Future empirical and theoretical studies will be required to clarify these issues.

We are very grateful to A. Clark and G. Gibson for helpful comments. This research has been supported by National Institutes of Health (NIH) grant RO1-GM36827 to M.L., and by graduate fellowships for A.F. funded by a National Science Foundation Training Grant in Genetic Mechanisms of Evolution and a NIH Training Grant in Developmental Biology.

LITERATURE CITED

Allendorf, F. W., F. M. Utter and B. P. May, 1975 Gene duplication within the family Salmonidae: II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations, pp. 415–432 in *Isozymes, Vol. IV: Genetics and Evolution*, edited by C. L. Markert. Academic Press, New York.

Amores, A., A. Force, Y.-L. Yan, L. Joly, C. Amemiya *et al.*, 1998 Zebrafish *hox* clusters and vertebrate genome evolution. Science **282:** 1711–1714.

Arnone, M. I., and E. H. Davidson, 1997 The hardwiring of development: organization and function of genomic regulatory sequences. Development **124:** 1851–1864.

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

Birky, C. W., Jr., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 6414–6418.

Brookfield, J. F. Y., 1997 Genetic redundancy. Adv. Genet. **36:** 137–155.

Christiansen, F. B., and O. Frydenberg, 1977 Selection-mutation balance for two nonallelic recessives producing an inferior double homozygote. Am. J. Hum. Genet. **29:** 195–207.

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

Clark, A. G., L. Wang and T. Hulleberb, 1995 Spontaneous mutation rate of modifiers of metabolism in *Drosophila*. Genetics **139:** 767–779.

DiLeone, R. J., L. B. Russell and D. M. Kingsley, 1998 An extensive 3′ regulatory region controls expression of *Bmp5* in specific anatomical structures of the mouse embryo. Genetics **148:** 401–408.

Donnelly, P., and S. Tavare, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

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

Fisher, R. A., 1935 The sheltering of lethals. Am. Nat. **69:** 446–455.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y.-L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.

Gerhart, J., and M. Kirschner, 1997 *Cells, Embryos, and Evolution*. Blackwell Science, Malden, MA.

Graf, J.-D., and H. R. Kobel, 1991 Genetics, pp. 19–34, in *Methods in Cell Biology, Vol. 36, Xenopus laevis: Practical Uses in Cell and Molecular Biology*, edited by B. K. Kay and H. B. Peng. Academic Press, New York.

Haldane, J. B. S., 1933 The part played by recurrent mutation in evolution. Am. Nat. **67:** 5–9.

Henikoff, S., E. A. Greene, S. Pietrokovski, P. Bork, T. K. Attwood *et al.*, 1997 Gene families: the taxonomy of protein paralogs and chimeras. Science **278:** 609–614.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Huang, J. D., D. H. Schwyter, J. M. Shirokawa and A. J. Courey, 1993 The interplay between multiple enhancer and silencer elements defines the pattern of decapentaplegic expression. Genes Dev. **7:** 694–704.

Hudson, R. R., 1991 Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. **7:** 1–44.

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

Jack, J., and Y. DeLotto, 1995 Structure and regulation of a complex locus: the *cut* gene of *Drosophila*. Genetics **139:** 1689–1700.

Kammandel, B., K. Chowdhury, A. Stoykova, S. Aparicio, S. Brenner *et al.*, 1999 Distinct *cis*-essential modules direct the time-space pattern of the *Pax6* gene activity. Dev. Biol. **205:** 79–97.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, United Kingdom.

Kimura, M., and J. L. King, 1979 Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA **76:** 2858–2861.

Kimura, M., and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kirchhamer, C. V., C.-H. Yuh and E. H. Davidson, 1996 Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc. Natl. Acad. Sci. USA **93:** 9322–9328.

Li, W.-H., 1980 Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics **95:** 237–258.

Li, W.-H., 1985 Accelerated evolution following gene duplication and its implication for the neutralist-selectionist controversy, pp. 333–352, in *Population Genetics and Molecular Evolution*, edited by T. Ohta and K. Aoki. Springer-Verlag, Berlin.

Lundin, L., 1993 Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics **16:** 1–9.

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

Nei, M., and A. K. Roychoudhury, 1973 Probability of fixation of nonfunctional genes at duplicate loci. Am. Nat. **107:** 362–372.

Normes, S., M. Clarkson, I. Mikkola, M. Pedersen, A. Bardsley *et al.*, 1998 Zebrafish contains two *Pax6* genes involved in eye development. Mech. Dev. **77:** 185–196.

Nowak, M. A., M. C. Boerlijst, J. Cooke and J. M. Smith, 1997 Evolution of genetic redundancy. Nature **388:** 167–170.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

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

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

Postlethwait, J. H., Y.-L. Yan, M. Gates, S. Horne, A. Amores *et al.*, 1998 Vertebrate genome evolution and the zebrafish gene map. Nat. Genet. **18:** 345–349.

Ramos-Onsins, S., and M. Aguade, 1998 Molecular evolution of the *Cecropin* multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Sidow, A., 1996 Gen(om)e duplications in the evolution of early vertebrates. Curr. Opin. Genet. Dev. **6:** 715–722.

Slusarski, D. C., C. K. Motsny and R. Holmgren, 1995 Mutations that alter the timing and pattern of *cubitus interruptus* gene expression in *Drosophila melanogaster*. Genetics **139:** 229–240.
Stoltzfus, A., 1999 On the possibility of constructive neutral evolution. J. Mol. Biol. **49:** 169–181.
Takahata, N., and T. Maruyama, 1979 Polymorphism and loss of duplicate gene expression: a theoretical study with application to tetraploid fish. Proc. Natl. Acad. Sci. USA **76:** 4521–4525.
Wagner, A., 1999 Redundant gene functions and natural selection. J. Evol. Biol. **12:** 1–16.
Walsh, J. B., 1995 How often do duplicated genes evolve new functions? Genetics **110:** 345–364.
Watterson, G. A., 1983 On the time for gene silencing at duplicate loci. Genetics **105:** 745–766.
Wendel, J. F., 1999 Genome evolution in polyploids. Plant Mol. Biol. (in press).
Westin, J., and M. Lardelli, 1997 Three novel *notch* genes in zebrafish: implications for vertebrate *Notch* gene evolution and function. Dev. Genes Evol. **207:** 51–63.
Xu, P.-X., X. Zhang, S. Heaney, A. Yoon, A. M. Michelson *et al.*, 1999 Regulation of *Pax6* expression is conserved between mice and flies. Development **126:** 383–395.

Communicating editor: A. G. Clark

APPENDIX

Consider a newly arisen mutant allele that is destined to become fixed in a population. In large populations, the time to fixation of an effectively neutral allele can be sufficiently long that all descendants of the single original mutant allele will have acquired secondary mutations by the time fixation of the entire lineage has occurred. Our concern here is with the probability that all descendants of a subfunctionalized allele destined to fixation have acquired nonfunctionalizing mutations by the time the gene pool consists entirely of descendants of the original subfunctionalized allele. To accomplish this, we take a gene genealogy approach, noting that the $2N$ genes present in a population at any point in time trace back to $2N - 1$ ancestral copies at some previous point in time, then to $2N - 2$ copies, etc., until the entire current population of genes coalesces to a single ancestral gene copy. Looking forward from the basal ancestral gene, the gene genealogy can be viewed as a series of bifurcations, with two branches emanating from the base of the genealogy, a third branch coming off one of these some time later, and so on.

Let $n = 2$ denote the first phase of the branching process, where there are only two independent branches, and $n = 3$ denote the next phase, in which a third branch joins the genealogy, etc. Conversion of all descendant genes in the lineage to a new mutational type occurs when independent mutations arise in the genealogy such that all descendants carry those mutations at the time of fixation of the lineage. For example, if both branches during period $n = 2$ independently acquire mutations, convergence has occurred because all subsequent descendants inherit those mutations. If only one of the two original branches acquires a mutation, conversion can still occur if, for example, the third branch in the genealogy coalesces with the branch initially containing the mutation (thereby gaining it by inheritance) and the nonmutant lineage acquires an independent mutation during $n = 3$. Mutational conversion can also occur if neither branch during period $n = 2$ acquires a mutation, but three or more mutations independently arise subsequently in appropriate positions in the gene genealogy.

To jointly account for the very large number of pathways by which mutational convergence can occur, we treat the branching process in the following way. Let $\pi_n(j)$ denote the probability that j of the n distinct terminal branches during interval n have acquired new mutations by the time phase n has elapsed, either by inheritance or by new mutation. The index j takes on values 0 to n, so $\pi_n(n)$ is equivalent to the probability that the lineage has been completely converted to the new mutational type by the end phase n. The quantity that we wish to estimate is $c = \pi_{2N}(2N)$, the ultimate probability that complete mutational conversion has occurred by the time the lineage has become fixed in the population. This quantity can be obtained recursively by defining a set of coefficients, $u_n(x, y)$, which denote the probability that x of y distinct branches acquire new mutations during interval n.

It follows that the probability that no mutations have arisen in the gene genealogy by the end of phase n is

$$\pi_n(0) = \pi_{n-1}(0) \cdot u_n(0, n), \tag{A1a}$$

which is simply the product of the probabilities that the genealogy is mutation-free entering the interval and the probability that none of the n unique branches in that interval acquire new mutations. The probability that one branch contains a mutation at the end of phase n is

$$\pi_n(1) = \pi_{n-1}(0) \cdot u_n(1, n) + \pi_{n-1}(1) \cdot \left(\frac{n-2}{n-1}\right) \cdot u_n(0, n-1). \tag{A1b}$$

Here, the first term is the probability that the lineage was mutation-free in the previous interval and that a single branch acquired a new mutation in interval n, whereas the second term is the joint probability that one mutation was present in the previous interval, that the branch added in this interval coalesces with any branch from the previous interval other than that carrying the mutation, and that no new mutation occurred in the current interval. The probabilities that two or more branches contain mutations at the end of phase n can be expressed by the general formula

$$\pi_n(j) = \pi_{n-1}(0) \cdot u_n(j, n) + \sum_{k=1}^{j-1} \pi_{n-1}(k) \cdot \left[\left(\frac{k}{n-1}\right) u_n(j-k-1, n-k-1) + \left(\frac{n-k-1}{n-1}\right) u_n(j-k, n-k)\right], \tag{A1c}$$

where $2 \leq j \leq n$. For each of the pairs of terms after the summation, the first member denotes the joint probability that the new branch added to the genealogy in interval n coalesces with a mutation-containing branch and that the necessary number of independent mutations occurs in mutation-free branches to yield j mutation-containing branches at the end of interval n. The second member of each pair is the alternative joint probability that the new branch coalesces with a mutation-free branch and that the necessary number of independent mutations occurs in it and/or other mutation-free branches to yield j mutation-containing branches at the end of interval n. The entire recursion initiates with $\pi_2(0) = u_2(0, n)$, $\pi_2(1) = u_2(1, n)$, and $\pi_2(2) = u_2(2, n)$.

To complete this exercise, we require expressions for coefficients of the form $u_n(x, y)$. Denoting t_n as the length of interval n in units of time, then $u_n(x, y)$ is simply the probability that during this interval x of y independent branches acquire a new mutation while $y - x$ do not. Letting μ denote the mutation rate, then assuming mutations are Poisson distributed, the joint probability of interest is

$$u_n(x, y)|t_n = \binom{y}{x}(1 - e^{-\mu t_n})^x \cdot e^{-(y-x)\mu t_n}. \quad \text{(A2)}$$

Further progress requires that we know something about the time interval t_n, but since we are dealing with effectively neutral mutations, this is straightforward, because t_n is simply the coalescence time during interval n. It is well known that under the Wright-Fisher model, coalescence times are distributed exponentially with expectation

$$\bar{t}_n = \frac{4N}{n(n-1)} \quad \text{(A3)}$$

(Kingman 1982; Hudson 1991; Donnelly and Tavare 1995). Integrating over the distribution of t_n, we obtain

$$\begin{aligned} u_n(x, y) &= \frac{1}{\bar{t}_n}\int_0^\infty [u_n(x, y)|t_n]e^{-t_n/\bar{t}_n}dt_n \\ &= \binom{y}{x}\sum_{i=0}^{x}\binom{x}{i}(-1)^{i+2}\frac{1}{(y - x + i)\mu\bar{t}_n + 1}. \end{aligned} \quad \text{(A4)}$$