Genetica **118:** 279–294, 2003.

Population-genetic models of the fates of duplicate genes

Bruce Walsh
Departments of Ecology and Evolutionary Biology, Molecular and Cellular Biology, and Plant Sciences, University of Arizona, Tucson, AZ 85721, USA (Phone: +1-520-621-1915; E-mail: jbwalsh@u.arizona.edu)

Key words: evolution of new gene function, gene families, gene silencing, neofunctionalization, pseudogenes, subfunctionalization

Abstract

The ultimate fate of a duplicated gene is that it either silenced through inactivating mutations or both copies are maintained by selection. This later fate can occur via neofunctionalization wherein one copy acquires a new function or by subfunctionalization wherein the original function of the gene is partitioned across both copies. The relative probabilities of these three different fates involve often very subtle iterations between of population size, mutation rate, and selection. All three of these fates are critical to the expansion and diversification of gene families.

The ultimate fate of a duplicated gene pair

The fate of duplicated genes is determined by the (often complex) interaction of three fundamental population-genetic forces: mutation, genetic drift and natural selection. These forces determine how often an initial duplication is fixed and decide the ultimate fate of any fixed duplication. The most obvious fate is that one of the duplicates is silenced, and the genome is littered with such pseudogenes. Some duplicates are inactivated immediately upon their formation, such as processed pseudogenes that derive from reverse-transcribed mRNAs or those that result from an incomplete duplication, and we ignore such genes here (see Walsh, 1985 for a treatment of the accumulation of processed pseudogenes). The other fate is that selection can maintain both duplicates. Indeed, most new genes are presumed to have arisen following a duplication event, wherein one copy maintains the original function while the other diverges and acquires a new function (we will refer to this fate as *neofunctionalization*). Alternatively, the duplicates can divergence in such a fashion that each takes over part of the function of the original gene, so that together both copies cover the original function but separately neither is sufficient (*subfunctionalization*). Under both neofunctionalization and subfunctionalization, selection maintains both copies in the face of mutational pressure for silencing.

This review of population genetic models of the fate of duplicated loci will largely follow along historical lines, as the historical development of duplication models reflects an increased sophistication in our understanding of their potential fates. Early models considered only the rate of silencing of established duplications. More recent models have considered the relative chances of whether an established duplication undergoes silencing versus neofunctionalization or subfunctionalization. Most recently, the process of fixation of initial duplication itself has been considered, because the copies may be silenced, neofunctionalized or subfunctionalized during the fixation process itself. We review recent developments in these areas and present some new results, and conclude with some general remarks on what all of the modeling has suggested to us about the evolutionary importance of gene duplication.

Population-genetic forces acting on duplicated gene

The apparent simplicity of a duplicate pair ultimately ending in one of three fates masks a number of very subtle interactions between population size, selection

and mutation that are only apparent from an investigation of the detailed population genetics. The role of any such analysis is to increase our insight into the relative importance of these various processes. For example, if we simply change the effective population size and no other parameters, what effect does this have on the ultimate fate of a duplication? As we will see, such a change can indeed have a rather dramatic effect.

Genetic drift, as measured by the effective population size N_e, is the simplest parameter in all the models. The time scale on which drift works scales with N_e, so that an allele is fixed more quickly in a smaller population (a single copy of a neutral allele destined to become fixed by drift takes, on average, $4N_e$ generations to do so). Likewise, changes in N_e change the relative effectiveness of any selection that may be operating, and at a sufficiently small N_e, the population dynamics of a selected allele is essentially the same as a neutral allele.

When considering mutation, both the mutation rate(s) and underlying mutational model need to be specified. In the simplest setting (which serves as a baseline for most models), it is assumed irreversible mutation occurs from functional copies to null (inactivated) copies. Models considering neofunctionalization or subfunctionalization require specification of the mutation rates from a functional copy to these different states, again under the assumption that once an allele mutates into one of these states, back mutation does not occur. For models examining neofunctionalization, it is assumed that at some (very low) rate functional alleles can mutate to an allele having a slightly different function (resulting in a selective advantage). Models allowing for subfunctionalization assume that the gene has two (or more) distinct functions. A fraction of mutations inactivate all functions, while others inactivate only one specific function, leaving the other(s) intact. Since the concern of the models examined here is the ultimate fate of the duplication pair, other mutations that change the sequence, but not the function, are simply ignored.

The final population-level process is selection. The default fitness model (unless otherwise specified) for null alleles is the *double recessive model*, in which individuals homozygous at both loci for null alleles have reduced fitness of $1 - s$ (s is often taken to be one, so that the double-recessive is lethal). Other fitness models allowing for partial dominance, wherein individuals carrying three (or even two) null alleles show reduced fitness, have also been considered. The true nature of the null-allele fitness function is very unclear. Li (1980) suggested that the double-recessive model is consistent with the data from the silencing of duplications in tetraploid fishes, as in many species, a large fraction (up to 70–80%) of duplicate loci have been silenced, and for every enzyme system examined, at least one species has lost a duplicate copy. Both these points suggest that genotypes carrying null alleles are likely weakly selected against in many cases. Conversely, Hughes and Hughes (1993) examined the synonymous and nonsynonymous rates of duplicate loci in the tetraploid frog *Xenopus laevis*, concluding that both copies of the duplicate genes are subjected to purifying selection, and hence the presence of a single null allele at a locus may also have fitness effects. One potentially ironic feature about mutations in duplicate genes is that missense mutations may be selected against, as these can potentially make a gene product that interferes with the normal gene function, while null mutations can be neutral as they will not generate such interference.

Some of the most subtle features of the various models examined below arise in large populations, and involve interactions between drift, mutation, and selection. When the combination of drift and effective population size is sufficiently small, each new mutation is either lost or fix before the next mutation appears. As population size increases, so does both the time to fixation under drift (which scales as N_e), increasing the chance that new mutations arise during the sojourn to fixation. Likewise, an increase in population size also increases the number of new mutations arising each generation (which scales as $N\mu$, the actual population size times the mutation rate). As a consequence, for sufficiently large populations, multiple alleles (e.g., functional, null, and alleles for new function) can all be segregating in the population. This significantly influences the relative fitnesses of these alleles compared with the much more straightforward situation where (at most) just two types of alleles are segregating. Likewise, a neutral allele drifting towards fixation can accumulate additional mutations during the time course of fixation, especially when this time is large relative to the mutation rate. Both these effects, changes in the relative fitness of an allele and conversion of an allele by mutation during fixation, are critical issues in the analysis of the fate of duplicate genes in large populations.

Finally, two other evolutionary forces, unequal crossing over and gene conversion, potentially play

important roles in the evolution of duplicate genes, especially when a number of copies are arranged in a tandem array or in tight linkage. These forces are important in the concerted evolution of duplicate copies, and the rich literature of population-genetic modeling of this process is reviewed in Ohta (1980, 1983, 1987) and Walsh (1987).

Models of strict gene silencing

Haldane (1933) was among the first to point out that one of the loci in a newly duplicated pair should be eliminated from the genome as it accrues mutations that silence it. He also made the important (and often overlooked) observation that duplication of a gene followed by silencing of the original (ancestral) copy can be a powerful force for changing the genetic map.

Motivated by Haldane's idea, Fisher (1935) presented the first population genetic model of the fate of duplicate genes. Fisher's context was the so-called *sheltering of lethals*, whereby an inactive copy (which would be lethal if the genotype was homozygous at both loci) can still be present at significant frequencies in a population when sheltered by the presence of a duplication. Fisher concluded that no locus would ever fix such an allele, but his model assumed both an infinite population size and reversible mutation (allowing the wild-type functional allele to be recovered from the null allele). In such a setting, the relative ratio of forward and back mutation rates, coupled with the amount of selection against genotypes bearing nulls, define an equilibrium frequency, typically resulting in the null alleles being rare at any given locus. A more detailed analysis of the selection-mutation balance in an infinite population was offered by Christansen and Frydenberg (1977).

Nei (1970, also Nei & Roychodhury, 1973) made the major contribution of introducing the effect of finite population size to Fisher's analysis. While still allowing for reversible mutations, Nei found that small populations had a very significant probability of being fixed for null alleles (lethal alleles in his terminology). As population size increases, back mutations became increasingly more important, so that null alleles were only expected to be fixed (more precisely, remain at frequency one for a substantial period of time) in moderate to small populations and loci can become unfixed by back mutation.

A landmark paper in the modeling of duplicate gene evolution is that of Bailey, Poulter and Stockwell (1978), who introduced a model of gene silencing with finite population size, double recessive lethal fitnesses for nulls and irreversible mutation. Biologically, this is a much more reasonable mutational model than that considered by Nei or Fisher in that one expects additional inactivating mutations to accumulate in a sequence long before the initial inactivating mutational site undergoes a reversion. (Gene conversion with a functional copy offers a possible route to reversion, but conversion tracks rarely cover an entire gene.) Bailey et al., were motivated by observations on several groups of tetraploid fish, most notably salmonids and catastomids. Electrophoretic analysis showed that even though the tetraplodization event for some groups may have been over 100 million years ago, duplicate pairs with two functional copies are still commonly seen. Thus, while many of the originally duplicated genes following the polyploidzation event became silent, many did not. This observation (see Ferris & Whitt, 1977, 1979, and Ferris, Portnoy & Whitt, 1979 for reviews of the data) sparked a flood of analyses of pure gene silencing models (e.g., Takahata & Maruyama, 1979; Li, 1980; Maruyama & Takahata, 1981; Watterson, 1983), and these served as the standard population-genetic models for the fate of duplicate genes until the mid-1990s. From their simulation studies, Bailey et al., concluded that the time for half the genes to become silenced scaled as $t_{1/2} \simeq 15N_e + \mu^{-3/4}$, where N_e is the effective population size and μ the mutation rate to null alleles. Based on this result, Bailey et al., concluded that their simple model of gene silencing was sufficient to account for the observed pattern of active and silenced genes in tetraploid fishes. They stressed the important point that the process of silencing a duplicate copy cannot start until disomic inheritance is reestablished following the polyploid event, further increasing the time for silencing. Simulation and analytic results from a number of authors (Nei & Roychodhury, 1973; Takahata & Maruyama, 1979; Li, 1980; Watterson, 1983) showed that Bailey et al.'s expression for $t_{1/2}$ was in error (as is detailed below).

A simple approximation related to the time to silencing was obtained by Nei and Roychodhury (1973), who used standard results from the neutral theory. Using the standard double-recessive model, they reasoned that for $4N_e\mu \ll 1$ null alleles behave in an essentially neutral fashion under the double recessive

fitness model, and obtained the probability that one null allele is fixed at one of the loci by generation t as

$$\Pr(\text{null fixed} \mid t \text{ generations}) \simeq 1 - e^{-\mu t}, \qquad (1a)$$

where μ is the null mutation rate. However, as Li (1980) points out, this result is for a single locus, and since one of the two loci in the pair (chosen at random) becomes inactivated, the actual pergeneration rate of silencing is double this, giving the fixation probability for a null at one locus in the duplicate pair as

$$\Pr(\text{null fixed} \mid t \text{ generations}) \simeq 1 - e^{-2\mu t}. \qquad (1b)$$

Both these expressions assume that the time to fixation of a successful null mutant (which scales as $4N_e$) is short relative to the time waiting for such a destined-to-be-fixed mutant to arise (which scales as $1/\mu$).

Using both simulation and analytic results, Takahata and Maruyama (1979) found that their expression for $t_{1/2}$ did not have the simple form suggested by Bailey et al., but rather was also a complex function of $N_e\mu$. Their analysis found that the rate of fixation of null alleles in tetraploid fish was much slower than predicted from theory. They also found that the fraction of polymorphic loci observed (those segregating significant frequencies of both active and null alleles) was much smaller than predicted under the double-recessive model, given the observed fraction of silenced loci. They thus rejected the double-recessive model as being sufficient to account for the observed maintenance pattern of duplicate loci seen in tetraploid fish. Takahata and Maruyama found that fixation times under the double-recessive fitness model were very similar to a strictly neutral model (no fitness effects whatsoever), unless $N_e s$ (s being the selection against the double-recessive) was extremely large (10^6 or greater). While the time to fixation increased with $N_e s$, it did so very weakly (the same conclusion was reached by Kimura & King, 1979). However, Li (1980) found that under other fitness models (such as individuals carrying three null alleles showing a slight reduction in fitness), the time to silencing was greatly increased over the strictly neutral model. In these cases, selection is clearly having an important effect in retarding silencing. Takahata and Maruyama thus concluded that there must be some partial dominance (individuals with three nulls showing reduced fitness) to account for the slower than expected rate of gene silencing in polyploid fish.

Maruyama and Takahata (1981) examined the effect of very tight linkage under the double-recessive model, finding that such linkage can greatly speed up the rate at which genes are silenced (relative to free recombination), especially in large populations. The reasoning for this effect can be seen by considering the case of complete linkage. Here, an *An* or *nB* (linked null *n* and active *A*, *B* alleles) gamete has no fitness effect under the double-recessive model (the active allele carried by each gamete masks any other nulls that might appear on the other gamete). Selection is thus very weak on these gametes, and the fixation times are closer to those expected under neutrality, resulting in a significant increase in the rate of silencing over loosely-linked loci (where, in a large population, the rate of silencing is much longer than under neutrality).

Li (1980) also examined the effects of linkage on the rate of silencing, concluding that tight linkage does not have a significant effect. However, Li's results and those of Maruyama and Takahata are actually rather similar in that the effect of tight linkage is to make a null allele behave more like a neutral allele, which results in a significant increase in the rate of silencing in a very large population when selection becomes increasingly important. Li also made the important observation that for small $N_e\mu$ values, the time to silencing (under the double-recessive fitness model) is half of that for a single loci neutral allele under irreversible mutation (as the actual mutation rate is twice that for a single locus, see Equation (1b)). However, as $N_e\mu$ increases, the expected silencing time under the double recessive model increases relative to the strict neutral model, eventually becoming considerably longer. The reason for this behavior is that when $N_e\mu$ is small, a null mutation will either be lost or fixed before a new null mutation arises at the other locus. Any null allele will thus still find an active copy in its background and hence will be selectively neutral, and since there are two loci at which null alleles can arise, the silencing rate is twice that for a single neutral locus. As $N_e\mu$ increases, so does the probability of null mutations arising at the other loci, in which cases a gamete bearing a null allele will have a relative fitness less than one and is (weakly) selected against. While the amount of selection (which depends of the frequencies of the null allele at both loci) can be very small, for a sufficiently large population size, the deterministic effects of selection can overpower the effects of drift for fixing null alleles.

The definitive treatment of silencing under the double null homozygote model was provided by Watterson (1983), who obtained an analytic approximation for the time to silencing that accounts for selection. If $1 - s$ denotes the relative fitness of the double null, then the expected time to silencing is

$$t \simeq N_e[\ln(2N_e s) - \Psi(2N_e\mu)], \tag{2a}$$

where Ψ denotes the digamma function. Since $\Psi(x)$ is a montonically increasing function, any increase in the mutation rate decreases the expected time to gene silencing. The asymptotic behavior of Ψ is as follows:

$$\Psi(x) \simeq \begin{cases} \ln x - 1/x & \text{for } x \gg 1, \\ 0 & \text{for } x = 1.462, \\ -4/x & \text{for } x \ll 1. \end{cases} \tag{2b}$$

Hence, for small null mutation rates, the time to fixation scales as least as slow as $1/2\mu$, which can greatly retard the time to fixation. However, for large values of $N_e\mu$, the decrease scales as the log of this product, so that a large increase in $N_e\mu$ results in only a small decrease in t. Likewise, since the effect of selection also scales logarithmically, large increases in $N_e s$ only result in small increases in the fixation time.

A very different model of gene silencing was proposed by Allendorf (1979). Under his *adaptive silencing* model, genotypes with four and three functional alleles are selected against (as is the double null homozygote). Under these fitness assumptions, selection drives one allele towards fixation, resulting in the rate of gene silencing increasing with population size (as opposed to decreasing under the mutational pressure model). Allendorf suggests that such a selection scheme can arise due to gene dosage issues following a polyplodization event. How realistic this model is remains uncertain, but if applicable, it is clearly restricted to the silencing of established duplications following a polyploid event. It is not a reasonable model for the silencing of a single gene duplication as the selection scheme would likely prevent the duplication from becoming established in the first place.

Models of pseudogenes versus neofunctionalization

Under the classic model of the fate of a pair of duplicate loci, while most acquire and fix null mutations at one member of the pair, a very small (but evolutionary extremely significant) fraction instead somehow acquire a new function (Ohno, 1970). In this setting, both the locus covering the ancestral function and neofunctionalized locus are subsequently maintained by selection. Walsh (1995) introduced a simple model for examining the relative probabilities of silencing versus neofunctionalization. The model assumes the standard double recessive model for null allele fitnesses, with μ being the mutation rate from functional to null alleles. Letting $\rho \ll 1$ denote the ratio of neofunctional to null mutations, then at rate $\rho\mu$ a functional allele mutates into a neofunctional allele with a slightly different function that conveys a selective advantage. The advantageous allele is assumed to have additive fitness, with a functional/neofunctional heterozygote having fitness $1 + s$ and homozygote fitness $1 + 2s$. Under the assumption that $N_e\mu \ll 1$, a mutation is either lost or fixed before additional mutations arise, and (recalling the above results from strict silencing models) null alleles behave as if they are neutral. Under this setting, the probabilities of neofunctionalization Pr(neo) and silencing Pr(sil) are determined by ρ and $S = 4N_e s$, the scaled selective advantage of a neofunctional allele, with

$$\begin{aligned} \Pr(\text{neo}) &= 1 - \Pr(\text{sil}) \\ &= \left(\frac{1 - e^{-S}}{\rho S} + 1\right)^{-1} \\ &\simeq \begin{cases} \rho & \text{for } S \ll 1, \\ S\rho & \text{for } S \gg 1,\ S\rho \ll 1, \\ 1 - 1/S\rho & \text{for } S\rho \gg 1. \end{cases} \end{aligned} \tag{3}$$

When the selective effect of the advantageous allele is sufficiently small ($S \ll 1$), the advantageous allele behaves as a neutral allele and the probability it becomes fixed (v.s. fixing a null allele) is a simple function of the relative mutation rates, with silencing being the typical fate as $\rho \ll 1$. The more interesting case is for strong selection on new advantageous alleles ($S \gg 1$), but advantageous alleles are still rare relative to the selection advantage, $S\rho \ll 1$. Here, null allele fixation (and hence silencing) is again the typical fate. Finally, if $S\rho \gg 1$, most duplicate pairs fix a neofunctional allele. Hence, larger population size favors neofunctionalization as this increases S.

Once the first advantageous allele is fixed at one of the loci, inactivating mutations continue to arise. If one of these becomes fixed, the neofunctional locus is converted to a pseudogene. However, unlike the initial situation where both loci have the same function, there is now selection against nulls at the neofunctional locus (as null/neofunctional heterozygotes have

fitness $1-s$). The same arguments leading to Equation (3) show that the probability of reinforcement (fixing a second advantageous allele as opposed to a null allele, and hence further increasing the selective barrier against fixing future nulls), is

$$\Pr(\text{2nd advantageous allele fixed}) \simeq \begin{cases} \rho & \text{for } S \ll 1, \\ 1 - e^{-S}/\rho & \text{for } S \gg 1. \end{cases} \quad (4)$$

Compared with the probability of neofunctionalization, note that reinforcement is far more likely (as selection scales as $1-e^{-S}$ in Equation (4) v.s. $1-1/S$ as in Equation (3)). Thus when $S \gg 1$, fixation of the first advantageous allele effectively locks the locus into a future of further functional divergence.

The analysis leading to Equation (3) assumes that the population size is not too large (i.e., $N_e\mu \ll 1$). When $N_e\mu$ is large, during the potential fixation of a new mutation, additional mutations, both null and neofunctional alleles, arise, and the presence of these additional mutations alters the fitness of null alleles. Consider a null allele segregating in the population and potentially on its way to fixation. For discussion, assume this null allele is at the ancestral locus *A* (as opposed to the daughter, or duplicate, locus *B*). The appearance of additional null mutations at locus *A* results in (at best) an extremely small increase in the probability of fixation of a null at *A*. However, null alleles arising at locus *B* now means that some rare individuals are double null homozygotes, resulting in a small amount of selection against null alleles at locus *A*. If the population size is extremely large, even this very small amount of selection can greatly retard the fixation of null alleles at either locus. A much more significant fitness effect occurs when a null allele is segregating and a neofunctional allele arises at either locus. In this case, the null allele is no longer neutral, but rather is at a selective disadvantage. Both of these effects result in Equation (3) being an underestimate, as the probability of fixation of the null allele is less than the neutral expectation in both cases.

Walsh (1995) offered a large population correction accounting for neofunctional alleles arising while nulls are otherwise drifting to fixation (as this is expected to be a much larger effect that the the selection against double-recessives). As a null allele is drifting towards fixation, on average a total of $4N_e^2$ copies of the functional allele exist over the entire time course to fixation. Thus, $4N_e^2\rho\mu$ advantageous mutations are expected to arise during this potential null fixation. Given that the probability of fixation of an advantageous allele introduced as a single copy is $S/2N$ (for $S \gg 1$), the probability that none of these advantageous alleles are fixed (so that the null indeed drifts to fixation) is

$$\left(1 - \frac{S}{2N}\right)^{4N_e^2\rho\mu} \simeq \exp(-\eta\rho S), \quad \text{where } \eta = 2N_e\mu.$$

If $\eta\rho S \gg 1$, then most paths counted in Equation (3) as having fixed null alleles actually fix advantageous alleles. Under this large population correction, the probability of neofunctionalization becomes

$$\Pr(\text{neo}) \simeq 1 - \frac{e^{-\eta\rho S}}{1+\rho S} \quad \text{for } S \gg 1. \quad (5)$$

For $\eta = 2N_e\mu \gg 1$, this results in a significant increase in the probability of neofunctionalization relative to Equation (3).

Analogous to the time to silencing is the more general *time to resolution* (Force et al., 1999), the time until the fate of the gene pair is resolved, be it neofunctionalized or a pseudogene. For a neutral allele, the expected waiting time between successful (destined to be fixed) mutations is just the reciprocal of the neutral mutation rate $(1/\mu)$, while the expected time to fixation of such a mutation is $4N_e$ generations. Hence, when $4N_e\mu \ll 1$, most of the time to resolution is spent waiting for a successful mutation, and we use this as an approximation for the resolution time. The per-generation rate at which destined-to-become-fixed null alleles arise is μ (the null mutation rate). Successful neofunctional alleles, on the other hand, arise at a per-generation rate of $(4N\mu\rho)(2s) = \mu\rho S$, the product of the total mutation rate and probability of fixation (assuming $N_e \simeq N$ and $4N_e s > 1$). Hence, the total rate at which successful mutations arise each generation is $\lambda = \mu(1+\rho S)$, and the distribution of the resolution time follows an exponential distribution with this parameter, giving the expected mean time to resolution as

$$E[t] = \frac{1}{\lambda} = \frac{\mu^{-1}}{(1+\rho S)}. \quad (6)$$

Likewise, the time $t_{1/2}$ for a 50% of the resolutions to have occurred is just $-\ln(0.5)/\lambda$, or 0.69 of the value in Equation (6). Note that the distribution of the resolution time conditioned on either an outcome of gene silencing or on an outcome of neofunctionalization, is the same. Hence Equation (6) is also the mean time to neofunctionalization (in those fraction

Pr(neo) of cases where this occurs) and the mean time to silencing in the remaining fraction of cases.

Models of pseudogenes versus subfunctionalization

The classic model for the fate of duplicate genes (silencing or new function) makes the implicit assumption that a single mutational event inactivates the entire gene. While this can be the case for mutations in the coding region, this mutational assumption is based on an overly simplistic view of the nature of the regulatory controlling elements. Genes often have multiple functions, and different (and often independent) regulatory elements to control expression in the tissues and/or environments corresponding to the different functions. Indeed, the different functions and their controlling elements partition the gene into a number of complementation classes. A mutation might thus completely disrupt the control element for one complementation class, but not effect any others. Such a mutation would be a null mutation for some functions of the gene, but not for others. This is the motivation for Force et al. (1999), who propose the duplication–degeneration–complementation (DDC) model for the fate of duplicate genes. Under this model, selection would preserve both copies if each fixes an inactivating mutant for a different complementation class of the ancestral function. Both copies together cover the ancestral function, yet inactivation of either copy would be deleterious as the other copy cannot fulfill all of the ancestral functions, a fate referred to as subfunctionalization. It is important to stress that while we have framed our introduction to functional complementation groups in the context of regulatory elements, the actual physical sites for these groups can also be alternate functional domains of the protein, splice site variants, etc. All that is required is mutation that can inactivate one function while not effecting any other.

Under the DDC model, the occurrence of mutations that only inactivate a specific subfunction is the key element that allows selection to maintain both copies. This model very nicely explains the confusing retention of duplicate loci in tetraploid fish, especially given the often extensive regulatory divergence that is also seen at these loci. An extremely important feature of this model is that by partitioning the ancestral function over the two copies, potential adaptive constrains on the evolution of the original gene may be softened, potentially allowing the gene to be subsequently neofunctionalized in some new direction. The fate of subfunctionalization thus increases evolutionary flexibility.

How often is subfunctionalization expected to occur? Force et al. (1999) and Lynch and Force (2000a) examined the relative probabilities of subfunctionalization versus silencing under the standard double-recessive (lethal) fitness model. They assume each gene has z independently mutable subfunctions, each with a null mutation rate of μ_r. A null mutation at one of the sites inactivates a particular subfunction without disturbing any of other the subfunctions. Likewise, other null alleles arise (e.g., in the coding region) at rate μ_c, and these mutants are assumed to inactivate all functions of the gene. Assuming that $N_e\mu_c \ll 1$, each null allele essentially behaves as a neutral, as any double null homozygotes are sufficiently rare to be ignored. The process of subfunctionalization is a two-phase process. First, one (or more) subfunctions must be inactivated (and fixed) in one locus. Second, the other locus must then fix a null allele at one of the subfunctions still active in the first locus. Once this occurs, both copies are needed for ancestral function and hence selection will maintain both copies. Further partitioning of the remaining shared subfunctions between loci is expected to continue following the initial subfunctionalization event.

Since both coding region nulls and subfunctional nulls behave neutrally (*provided* a functional site is present at the other locus), the probability that a null is fixed in a subfunctional region first (as opposed to the coding region) is just the ratio of the total mutation rate in the subfunction region ($z\mu_r$) to the total mutation rate ($z\mu_r + \mu_c$), giving the probability of a successful first phase of subfunctionalization as

$$\Pr(\text{first phase}) = \frac{z\mu_r}{\mu_c + z\mu_r}. \tag{7a}$$

Let *A* denote the locus that fixes the first subfunctional null, and *B* the other locus. Locus *A* now has $z - 1$ active subfunctional sites. Since this missing subfunction is assumed vital, selection will quickly remove any null mutations in *B* that occur in either this subfunctional region or in the coding region. However, mutations that inactivate of any of the remaining $z - 1$ subfunctional sites active in both genes behave as a neutral. Subfunctionalization occurs immediately if an inactivating mutation in one of these remaining common subfunctional regions is fixed for locus *B*. The

probability of this event is

$$\Pr(\text{sub in 2 events}) = P_{S,2} = \left(\frac{z\mu_r}{\mu_c + z\mu_r}\right)\left(\frac{(z-1)\mu_r}{\mu_c + 2(z-1)\mu_r}\right). \quad (7b)$$

The first term corresponds to the probability of losing a subfunctional site at locus *A* and the second term the probability of fixing a null mutation in a different subfunction at locus *B*. This second term follows since permissible null mutations (those not quickly removed by selection) are those that inactivate of the coding region or any of the remaining $z-1$ control regions in locus *A* or in the $z-1$ control elements in locus *A*, for a total rate of $\mu_c + 2(z-1)\mu_r$. Of this total rate, $(z-1)\mu_r$ is the rate of inactivation of one of the control regions in locus *B*, which results in subfunctionalization. Two other outcomes are possible: fixing a null mutation in the coding region of *A* (silencing the gene) or fixing a null mutant in another subfunctional region of locus *A*. In the latter case, subfunctionalization can still occur, provided no coding regions nulls are fixed in *A* and a new subfunction null is eventually fixed in locus *B*. In general, subfunctionalization may take $j \leq z$ steps, with $j-1$ elements inactivated in locus *A* before an independent element is inactivated in *B*, after which point selection maintains both loci. Following the same logic leading to Equation (7b), the probability of subfunctionalization after *j* fixed mutations is

$$\Pr(\text{sub in } j \text{ events}) = P_{S,j} = \left(\frac{z\mu_r}{\mu_c + z\mu_r}\right)\prod_{k=0}^{j-2}\left(\frac{(z-k-1)\mu_r}{\mu_c + 2(z-k-1)\mu_r}\right). \quad (7c)$$

The overall probability of subfunctionalization is just the sum over all possible events,

$$P_s = \sum_{j=2}^{z} P_{S,j}. \quad (7d)$$

Figures 1 and 2 plot P_s for $z = 2$–5 as a function of $\gamma = \mu_r/\mu_c$ (the ratio of subfunctional nulls to coding region nulls). Increasing the number of subfunctions *z* or increasing γ (increasing μ_r relative to μ_c) increases the probability of subfunctionalization. Force et al., note that if the total mutation rate over the subfunctions is greater than four times the null rate in the coding region ($z\mu_r > 4\mu_c$), then the probability of subfunctionalization exceeds 50%. For very

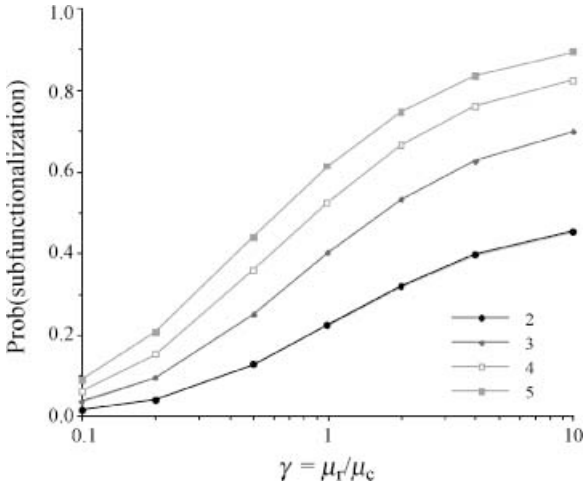

Figure 1. The probability P_s of subfunctionalization (v.s. gene silencing), as a function of $\gamma = \mu_r/\mu_c$ (the ratio of the subfunctional null to coding region null mutation rates) and the number of subfunctional sites (here *z* ranges from 2 to 5).

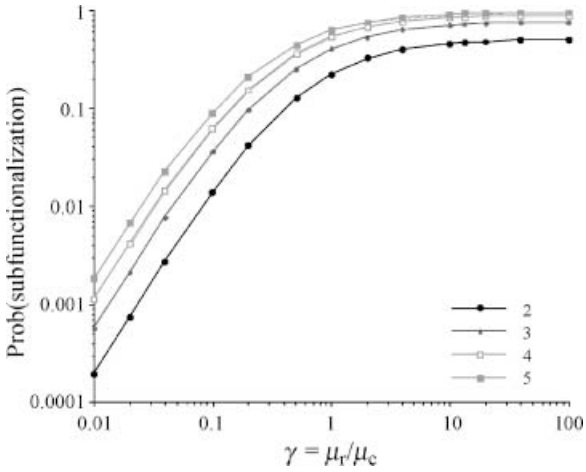

Figure 2. The probability P_s of subfunctionalization (on a log scale) for a broader range of $\gamma = \mu_r/\mu_c$ values for two to five subfunctions.

large values of γ (i.e., $\mu_r \gg \mu_c$), the probability of subfunctionalization approaches

$$P_s \to 1 - \left(\frac{1}{2}\right)^{z-1} \quad \text{as } \gamma \to \infty. \quad (7e)$$

For example for $z = 2$, P_s approaches a limiting value of 1/2 when the mutation rates for the subfunctions are much greater than for the coding region. This makes sense in that, following fixation of a null in the first subfunctional, there is an equal chance that the second subfunction is fixed for a null in the same gene (silencing) or in the other copy (subfunctionalization).

Force et al., showed, for those genes destined to become subfunctionalized, that the mean time to

resolution is

$$t_s = \sum_{i=2}^{z} \frac{t_{s,i}\ P_{S,i}}{P_s}, \quad \text{where}$$

$$t_{s,i} = \frac{1}{\mu_r}\left(\frac{1}{2z} + \sum_{j=1}^{i-1} \frac{1}{z-j}\right). \tag{8}$$

Note that the time scales as the reciprocal of the subfunctional mutation rate μ_r. Force et al., note that for $\mu_r = 10^{-7}$, then t_s does not exceed 12.5 million years, and is on the order of 4 million years or less when $z > 5$.

Equations (7) and (8) were obtained when the product of the effective population size and mutation rates are much less than one. Two factors come into play when this product ($N_e\mu$) exceeds one. First, there is some selection against nulls, as although double null homozygotes are still very rare, the population size can be sufficiently large for even this small amount of selection to be important. However, this effect is usually a very small unless the effective population size is extremely large (Equation (2) shows that the effect of selection scales as $\ln 2N_e s$). A much more significant effect of increasing effective population size was detected by Lynch and Force (2000a), a process they refer to as *mutational conversion*. When $N_e(\mu_c + \mu_r)$ is large, additional null mutations (at other sites) are expected to arise during the course of fixing a particular subfunctional null in the population. As a result, in large populations, by the time a null for a subfunctional site is fixed, it has also acquired a number of other nulls, so that the original subfunctional allele has become silenced (converted to a silenced gene) by these secondary mutations. Mutational conversion results in the entire process of subfunctionalization grinding to a halt in sufficiently large population. Since a neutral allele takes (on average) $4N_e$ generations to drift to fixation, the probability that alleles which descend from an original mutation *have not* experienced a null mutation in the coding region is $\exp(-4N_e\mu_c)$, which is 0.02 for $N_e\mu_c = 1$ and 10^{-7} for $N_e\mu_c = 4$. Lynch and Force show that tight linkage between the duplicates increases the probability of subfunctionalization relative to unlinked loci, but this effect ultimately is small and does not allow for subfunctionalization in large populations.

A unified (small population) model

It is straightforward to combine the models of Walsh (1995) and Force et al. (1999) into a single (small population) model that allows for all three possible fates, silencing, neofunctionalization, subfunctionalization. We present the analysis for two subfunctions ($z = 2$), as extension to larger values of z follows in an obvious fashion. Figure 3 defines the various per-generation rates for transition between the various states. Under the small population assumption that nulls and subfunctional alleles behave as neutral alleles under permissible conditions and that mutational conversion can be ignored, results from the previously-discussed models give these rates as

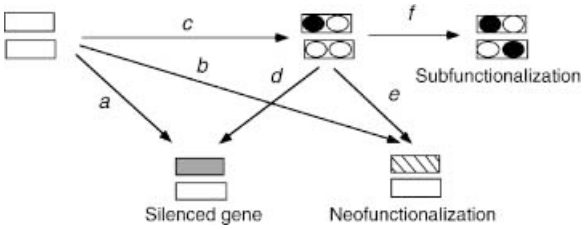

Figure 3. Combined fates of pseudogene versus neofunctionalization versus subfunctionalization. We assume $z = 2$ subfunctions (circles, which are open when active, filled for inactive). For an initial duplication fixed for two functional copies, three initial fates are possible. Two of these fates are terminal events: with rate a it can fix a null allele and become a pseudogene, and with rate b it can fix a neofunctionalizing allele. Otherwise, at rate c it can fix a mutation inactivating one of the subfunctions, which can result in three possible resolutions. With rate f the other copy loses the other subfunction and subfunctionalization occurs, while with rate d a null allele is fixed and at rate e an neofunctional allele is fixed.

$$a = \mu_c, \qquad b = \mu_c\rho S, \qquad c = 2\mu_r,$$
$$d = \mu_c + \mu_r, \qquad e = \mu_c\rho S, \qquad f = \mu_r, \tag{9a}$$

where we now define $\rho = \mu_n/\mu_c$ as the ratio of the neofunctional mutation rate (μ_n) divided by the coding region null mutation rate and $\gamma = \mu_r/\mu_c$ and $S = 4N_e s$ (as before). We can rescale these rates by dividing through by the coding region mutation rate μ_c to give relative rates of

$$a = 1, \qquad b = \rho S, \qquad c = 2\gamma,$$
$$d = 1 + \gamma, \qquad e = \rho S, \qquad f = \gamma. \tag{9b}$$

Note that only two parameters, γ and ρS, are required to prescribe the probabilities of the various fates (which are functions of ratios of the rates). An interesting possibility (easy to incorporate by defining $e = \mu_c\rho^* S^* > b = \mu_c\rho S$) is that the removal of some pleotropic constraint due to inactivation of one of the subfunctions increases the chance of a neofunctional

mutation ($\rho^* > \rho$) and/or its selective advantage ($s^* > s$).

Following the same logic that leads to Equation (7b), the probability that subfunctionalization is the ultimate fate is

$$\begin{aligned}&\Pr(\text{subfunctionalization})\\ &\quad=\left(\frac{c}{a+b+c}\right)\left(\frac{f}{d+e+f}\right)\\ &\quad=\frac{2\gamma^2}{(1+2\gamma+\rho S)^2}.\end{aligned}\tag{10a}$$

Note that this probability is a decreasing function of ρS, so that any possibility of neofunctionalization ($\rho S > 0$) decreases the chance of subfunctionalization. However, unless ρS is of order one or larger, neofunctionalization has only a negligible effect on decreasing the chance of subfunctionalization.

Similarly, a gene may be either neofunctionalized in the first mutational event or following a mutation inactivating one of the subfunctions (Figure 3), giving

$$\begin{aligned}&\Pr(\text{neofunctionalization})\\ &\quad=\frac{1}{a+b+c}\left[b+\left(\frac{ce}{d+e+f}\right)\right]\\ &\quad=\frac{\rho S(1+4\gamma+\rho S)}{(1+2\gamma+\rho S)^2}.\end{aligned}\tag{10b}$$

One can easily show that this probability is a decreasing function of γ. Hence, the possibility of subfunctionalization ($\gamma > 0$) lowers the probability of neofunctionalization. Figure 4 shows the behavior of these probabilities. The probability that both genes

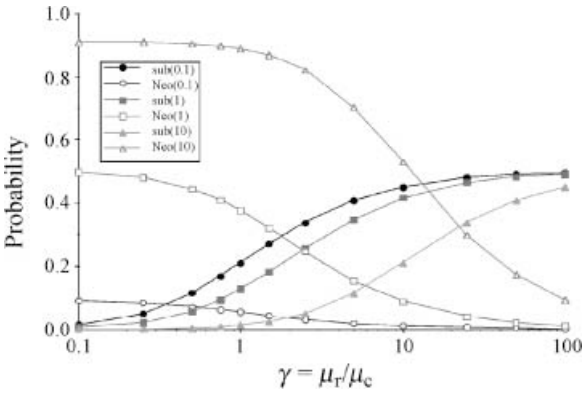

Figure 4. The probabilities of neofunctionalization (open points) and subfunctionalization (solid points) as a function of $\gamma = \mu_r/\mu_c$ and $\rho S = 0.1$ (circles), 1 (squares), and 10 (triangles).

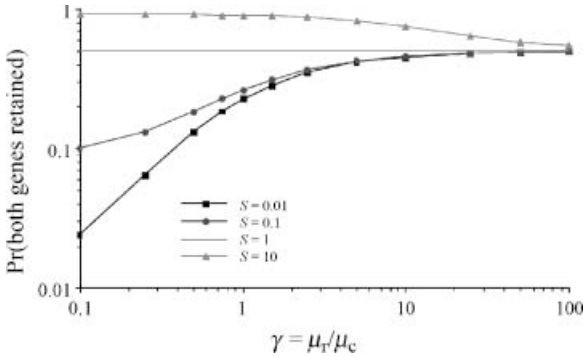

Figure 5. Probability of retaining both loci, as a function of $\gamma = \mu_r/\mu_c$ and ρS.

are maintained is given by the sum of Equations (10a) and (10b),

$$\Pr(\text{retention}) = 1 - \frac{1+4\gamma+2\gamma^2+\rho S}{(1+2\gamma+\rho S)^2}\tag{11}$$

which is plotted in Figure 5. By taking derivatives with respect to the two parameters (γ and ρS), one can easily show that the probability of retention is an increasing function of ρS. However, the sign of the derivative with respect to γ equals sign $(1-\rho S)$. Hence, for $\rho S < 1$, the probability of retention is an increasing function of γ, while when $\rho S > 1$ increases in γ *decreases* the probability of retention. The reason for this behavior follows from a consideration of the fate of a duplicate when γ is very large, in which case the probability of subfunctionalization approaches its limiting value given by Equation (7e) (1/2 for $z = 2$). For smaller values of γ, when $\rho S > 1$, neofunctionalization is more important than silencing and the probability of retention exceeds 1/2. However, as γ increases and subfunctionalization comes to dominate the system, the probability of retention decreases to the limiting subfunctionalization value.

Finally, the mean time to resolution follows from a simple conditioning argument. The expected time to fix the first mutation is the reciprocal of the total rates, $1/(a+b+c)$. With probability $(a+b)/(a+b+c)$ this first fixed mutation results in a resolution (silencing or neofunctionalization). With probability $c/(a+b+c)$ a second mutation is required for resolution, and the mean time after the first mutation is fixed until the second is fixed is $1/(d+e+f)$. Putting these together, the mean time is

$$\begin{aligned}E[t] \;=\; &\left(\frac{a+b}{a+b+c}\right)\left(\frac{1}{a+b+c}\right)+\\ &+\left(\frac{c}{a+b+c}\right)\left[\frac{1}{a+b+c}+\frac{1}{d+e+f}\right]\end{aligned}$$

$$= \frac{c+d+e+f}{(a+b+c)(d+e+f)}$$
$$= \mu_c^{-1}\frac{1+4\gamma+\rho S}{(1+2\gamma+\rho S)^2}. \quad (12)$$

The resolution time thus scales as the reciprocal of the coding region mutation rate and is a decreasing function of both γ and ρS.

Models for the establishment of the initial duplication

Up to this point, all models have made the assumption that the population starts out as being fixed for two fully-functional duplicates. Certainly this is true immediately following a polyplodization event, but more generally when a single locus or some larger chromosomal segment is duplicated, the duplication must first become fixed in the population before we can start to worry about its ultimate fate.

There are two key issues here. First, what forces are involved in fixing a duplication and, second, does this fixation process alter the potential fate of the duplicated gene? Clearly if the duplication event inactivates the gene, its fixation results in a fixed silenced duplication. Likewise, the dynamics discussed above for the three competing processes (silencing, neofunctionalization, subfunctionalization) also occur when an initial duplication is undergoing its sojourn to fixation, and in many cases there is a resolution of the outcome as a direct consequence of the fixation process.

While it is generally thought that most new duplications simply drift to fixation, the situation can be considerably more complicated in that selection can have a significant influence on fixation. Clark (1994) showed that new duplications can have a slight selective advantage as they mask the effects of deleterious mutations at the functional loci. Given that the strength of any such selection is on the order of the mutation rate, this effect should not be very significant except in very large populations. A much more significant fitness effect of new duplications was examined by Spofford (1969). Building on the suggestion of Partridge and Giles (1963), Spofford examined a locus under overdominant selection, so that the heterozygote has the highest fitness. In this case, a duplication is immediately favored by selection. To see this, suppose the genotype *AB* has the highest fitness, and the *AA* and *BB* homozygotes have equal (and lower) fitness. At the newly duplicated locus (say for allele *A*), the marginal fitness of individuals carrying a copy of *A* is greater than the marginal fitness for individuals missing the duplication (let *O* denote the 'allele' that is missing the new duplication), as the fitness BB/A is greater than the fitness of BB/O. Fixation of the *A*-bearing duplication then drives *B* to fixation at the original locus, and the net result is that all individuals in the population are $AABB$, increasing population fitness relative to the single-locus case. In this setting, the duplication of a pre-existing allele greatly increases the probability that the duplication is fixed.

A similar situation can occur for neofunctional alleles. Let *f* denote a functional allele, *n* a neofunctional allele. Obviously, those mutations that not only give a new (advantageous) function but also retain the original function will be driven by selection to fixation even at single loci. Hence, the neofunctional alleles of interest are those where allele *n* has a fitness advantage of $1+s$ when heterozygous with a functional allele (i.e., *nf*), but are deleterious or lethal as an *nn* homozygote as they cannot cover all functions of the original allele. When *nn* is lethal, the equilibrium frequencies of the neofunctional and functional alleles at a single locus in an infinite population (without mutation) are easily obtained. Given that the genotypes *nn:fn:ff* have relative fitnesses of $0{:}1+s{:}1$, it immediately follows that:

$$\widehat{p}_n = \frac{s}{1+2s} \simeq s, \quad \text{and}$$
$$\widehat{p}_f = \frac{1+s}{1+2s} \simeq 1-s. \quad (13a)$$

More generally, both functional and neofunctional alleles are being mutated to null alleles at rate μ_c, and this influences the equilibrium frequencies. When mutation is strong relative to selection ($\mu_c > s^2$), then $\widehat{p}_n \simeq 0$, and neofunctional alleles are essentially absent at the single locus. In such cases, a new duplication will almost certainly be formed with a functional allele (the frequency of null alleles scaling as $\sqrt{\mu_c}$) and will have to subsequently acquire a neofunctional allele if neofunctionalization is to occur. However, if selection is strong relative to mutation ($\mu_c < s^2$), then

$$\widehat{p}_n \simeq \frac{s^2-\mu_c(1+s)^2}{s(1+2s)}$$
$$\simeq s(1-\mu_c) \quad \text{when } s \gg \mu_c \quad (13b)$$

as obtained by Lynch et al. (2001). Here, a reasonable fraction of existing alleles will be neofunctional, and an (unlinked) duplication can start off as a neofunctional allele. Simulation studies by Lynch et al., show

that an additional condition is required for Equation (13b) to hold, namely that the population size is sufficiently large, otherwise the frequency of neofunctional alleles in the population will be significantly below the deterministic value given by Equation (13b). In particular, unless $Ns^2 > 4$, neofunctional alleles will have negligible frequencies and a new duplication will almost always start with a functional allele.

Suppose that Equation (13b) does indeed hold. When the duplication is formed from a neofunctional allele (the duplication is segregating *n* and *O*, i.e., no duplication), then ignoring nulls, the two-locus fitnesses are

New duplication	Ancestral locus		
	ff	*fn*	*nn*
nn	$1+2s$	$1+2s$	0
nO	$1+s$	$1+2s$	0
OO	1	$1+s$	0

These fitnesses show that the neofunctional allele at the duplication has an immediate selection advantage over the no-duplication allele (*O*). Since the selective advantage is of order *s*, for a large population the probability that the *n* allele at the new duplication is fixed is of order $2s$. Assuming that the duplication occurs by randomly choosing an existing allele, the probability that the initial duplication involves *n* is also of order *s* (Equation (13b)), giving the total probability of a duplication both starting with, and subsequently fixing, a neofunctional allele as being of order $2s^2$. Conversely, if a functional allele is chosen to start the duplication, the fitnesses now become

New duplication	Ancestral locus		
	ff	*fn*	*nn*
ff	1	$1+s$	$1+2s$
fO	1	$1+s$	$1+2s$
OO	1	$1+s$	0

Notice that the presence of the *f* allele at the duplication transforms the nature of selection on the ancestral locus frombalancing for both *f* and *n* to directional for *n*. The *f* allele at the new duplication has a fitness advantage of freq $(nn)(1+2s)$ through being associated with the *nn* genotype, giving a selection coefficient on the order of s^2 (assuming Equation (13b) holds). Since the frequency of an *f* allele is $1-s^2 \simeq 1$, the probability that the new duplication fixes *f* and drives the neofunctional allele towards fixation at the ancestral locus is also of order $2s^2$. With *f* fixed at the new duplication, directional selection will then fix *n* at the ancestral locus. While drift can still prevent this second fixation from occurring, for large populations the chance of *n* not being fixed is very small, as it does not start from a single copy, but rather from some significantly higher frequency. See Lynch et al. (2001) for a more careful treatment.

A final setting where selection can foster the fixation of a duplication is if a neofunctional mutation occurs in the segregating duplication on its way to either loss or fixation. When a new duplication arises, it is assumed to drift to fixation with a probability equal to its initial starting frequency $1/2N$, giving the probability of loss for a new (neutral) duplication as $1-1/2N$. However, even when a duplication is destined to become lost, it can still drift around in the population for some time, potentially allowing it to acquire a rare neofunctional mutation. In this case, the new duplication is now under selection, which can then drive it to fixation, and in the process immediately neofunctionalize the duplicate pair. We can examine the potential impact of such rescuing of otherwise destined-to-best lost duplications by using very similar arguments to those applied for Walsh's (1995) large population correction. Accounting for the total number of copies of the new duplication that appear during the sojourn of the duplication from its initial appearance until it would otherwise be lost by drift, the expected number of advantageous (neofunctional) alleles can be obtained. Weighting this by the probability of fixation, it is found that the fraction of new duplications assumed lost by drift that are actually fixed by acquiring an advantageous mutant is roughly $1-\mathrm{e}^{-\mu\rho S} \simeq \mu\rho S$ (for small values of $\mu\rho S$). Since the probability a new duplication is lost is $1-1/2N$, the probability that neofunctionalization occurs via this pathway is

$$\left(1-\frac{1}{2N}\right)\mu\rho S \simeq \mu\rho S.$$

Several of the above points were synthesized in the recent work of Lynch et al. (2001) to provide a working framework for a general theory on the fate of newly arising duplications. These authors focus on three summary statistics for a newly arising duplication: (i) Θ, $2N$ times the probability that the new duplication is permanently preserved, (ii) Γ, $2N$ times the probability that both the daughter and ancestral locus are jointly preserved, either by subfunctionalization or neofunctionalization, and (iii) Δ, $2N$ times the

probability that the ancestral locus (or some subfunction of it) is reassigned to a new genomic location. These probabilities are scaled by $2N$, as the probability of fixation of a neutral allele introduced as a single copy is $1/2N$, so that scaled values equal to one are equal to the neutral values, values greater than one are larger than a neutral, and less than one are smaller than a neutral. The parameters Γ and Δ are indicators of the genome-wide implications of continual dupication, with Γ a measure of the expansion of gene number and Δ a measure of functional shuffling over genomic locations. Further, since the expected (per generation) rates at which these events occur is $2N\mu$ times the probability of the event, where μ is the duplication rate. Hence $\mu\Theta$, $\mu\Delta$, and $\mu\Gamma$ are the expected per-generation rates for these event. For example, the probability (after t generations) of no genomic expansion involving a particular locus is just $\exp(-\mu t\Gamma)$.

One general conclusion from the simulations and analysis of Lynch et al., follows from consideration of the strict silencing model. In this case, the probability of fixation for a new duplication is on the order of $1/2N$, while one of the two copies (either the original or the new duplication, chosen at random) ultimately becomes silenced. Thus, under strict silencing the probability that a new duplication is permanently preserved (as opposed to first being fixed and subsequently being silenced) is $(1/2)1/2N = 1/4N$, giving $\Theta = 1/2$. Unless there is active selection against a new duplication, this is the lower limit for the probability of fixation and preservation. Likewise, the probability that the ancestral gene function changes genomic location (here, the new duplication becomes fixed, and the ancestral copy fixes a null allele) is $(1/2)1/2N = 1/4N$ or $\Delta = 1/2$. These values under strict silencing serve as a useful benchmark for the effects of subfunctionalization and neofunctionalization. Under strict silencing, there is no permanent expansion of gene number, so that $\Gamma = 0$, while the rate of genomic shuffling of gene function is on the order of half the duplication rate. Thus the probability that no genomic shuffling for a particular locus will have occured after t generations is $\exp(-\mu t\Delta) = \exp(-\mu t/2)$.

Now suppose that subfunctionalization can occur. If $P_{\mathrm{sil},A}$ denotes the probability of silencing at the ancestral locus and P_{sub} the probability of subfunctionalization, then it follows that:

$$\Theta = \Delta = 2N(P_{\mathrm{sil},A} + P_{\mathrm{sub}}), \quad \text{and}$$
$$\Gamma = 2NP_{\mathrm{sub}}. \tag{14a}$$

These probabilities follow in that a new duplication is preserved if either the original locus is silenced or if subfunctionalization occurs. The P_{sub} term in Δ (the scaled probability of a map change) occurs because the subfunctionalization event partitions the original function across both loci, resulting in a genomic dispersion of the original function. For small to medium population sizes, simulation studies by Lynch et al., show that the probability of fixation of the initial duplication is very close to its initial frequency, $1/2N$. Let $P'_{\mathrm{sil},A}$ and P'_{sub} denote the probabilities of these events, conditioned on the duplication becoming fixed. If silencing occurs, it effects the original and duplicate loci with equal probability, implying $P'_{\mathrm{sil},A} = (1/2)(1 - P'_{\mathrm{sub}})$, giving

$$\Theta = \Delta = 2N\frac{1}{2N}\left(\frac{1 - P'_{\mathrm{sub}}}{2} + P'_{\mathrm{sub}}\right) = \frac{1 + P'_{\mathrm{sub}}}{2} \tag{14b}$$

and

$$\gamma = 2N\frac{1}{2N}P'_{\mathrm{sub}} = P'_{\mathrm{sub}}, \tag{14c}$$

where the conditional probability of subfunctionalization for small populations is given by Equation (7d). As the population becomes sufficiently large, mutational conversion will essentially prevent subfunctionalization from occurring, so that $P'_{\mathrm{sub}} \to 0$ for very large N, so that $\Theta = \Delta \to 1/2$ and $\Gamma \to 0$.

Another interesting feature for large populations was observed in the simulations when the duplication was completely linked to the ancestor. In such cases, one can think of the duplication as being a two-copy 'allele' and the unduplicated region as the single-copy allele. In very large populations, the two-copy allele is at a very slight advantage over single-copy alleles, as the later can be inactivated by a single mutation. This results in the probability of fixation of the duplication approaching $1/N$, giving

$$\Theta = \Delta = 2N\frac{1}{N}\frac{1}{2} = 1. \tag{15a}$$

For a pair of unlinked loci in a large population, there is no large-population selective advantage to a new duplication, so that the probability of fixation of a new duplication is $1/2N$. Again, in a large population, the probability of subfunctionalization approaches zero due to mutational conversion of a duplication on its sojourn to fixation. Hence, for unlinked loci, in a large population

$$\Theta = \Delta = 2N\frac{1}{2N}\frac{1}{2} = \frac{1}{2}. \tag{15b}$$

Thus, in large populations, the probability of fixation for a completely linked duplication is twice that of an unlinked one (as also noted by Li, 1980 and Maruyama & Takahata, 1981). In either event, the actual probability of fixation is still very small (scaling as $1/N$). Notice that in large populations, $\Gamma \to 0$, so that the rate of genomic expansion due to the preservation of newly duplicated genes is very small, but that the rate at which functions becomes shuffled over the genome remains on the order of the duplication mutation rate (as Δ ranges from 0.5 for unlinked duplications to 1.0 for completely linked duplications). A final complication observed in the simulations of Lynch et al., for very large populations (on the order of 10^7) is that while Δ (the scaled map change probability) approaches one for completely linked loci, it decreases below 0.5 for unlinked loci. There are two potential sources for this decrease. First, at very large populations, there can be weak selection against a null, decreasing the probability of fixation below the neutral value. A second and a more likely explanation is mutational conversion, with the descendants of the new duplication being silenced by mutation while the locus is drifting toward fixation.

Turning to silencing versus neofunctionalization, the scaled probability of preservation of a new duplication now becomes

$$\Theta = 2N(P_{\text{neo},A} + P_{\text{neo},D} + P_{\text{sil},A}), \tag{16a}$$

where $P_{\text{neo},x}$ denotes the probability of fixing a neofunctional allele for the ancestral ($x = A$) and duplicate ($x = D$) loci. If a neofunctionalizing allele is not fixed at either locus, the new duplication is still preserved if the ancestral copy is silenced, yielding Equation (16a). Likewise, the scaled probability Γ of genome expansion is just $2N$ times the probability of neofunctionalization, while the scaled probability of a map change becomes

$$\Delta = 2N(P_{\text{neo},A} + P_{\text{sil},A}). \tag{16b}$$

For a small population, one would expect neofunctional alleles to be essentially absent from the population, and the probability of fixation of a new duplication is $1/2N$, the neutral value. Conditioned on the duplication being fixed for a functional allele, Equation (3) gives the probability of neofunctionalization, which we denote by β. Since both loci have an equal chance of neofunctionalization,

$$\begin{aligned} P_{\text{neo},A} &= P_{\text{neo},D} = (\beta/2)1/2N, \quad \text{and} \\ P_{\text{sil},A} &= P_{\text{sil},D} = (1-\beta)(1/2)1/2N \end{aligned} \tag{16c}$$

implying

$$\begin{aligned} \theta &= \beta + (1-\beta)/2 = \tfrac{1}{2}(1+\beta), \quad \text{and} \\ \Delta &= \beta/2 + (1-\beta)/2 = 1/2. \end{aligned} \tag{16d}$$

Notice that the probability of a change in map position is not affected by the probability of neofunctionalization. This occurs because, regardless of whether silencing or neofunctionalization happens, with probability $1/2$ it involves the original gene, and hence a change in the genomic location of the function. Likewise, note in the small populations that Θ and Γ both approach a limit of 1 as the probability of neofunctionalization (conditioned on starting with two functional, fixed, duplications) $\beta \to 1$. Thus, none of these statistics exceeds the neutral expectation in small populations, even when β is close to one.

For larger populations, the probability of neofunctionalization increases dramatically, as selection can drive the duplication to fixation through the three pathways detailed above, all of which give fixation probabilities that approach a constant (twice the selection coefficient) instead of decreasing as $1/2N$ as is the case for a neutral gene. Hence, as N increases, we expect all three scaled parameters (Θ, Δ, Γ) to scale with N (i.e., for large N, they behave like aN). The first two selection pathways require a nontrivial frequency of neofunctional alleles segregating in the population. In this case, either the duplication starts with a neofunctional allele which is them fixed (with probability scaling as $2s^2$) or the duplication starts with a functional allele, which subsequently drives a neofunctional allele to fixation at the ancestral locus, again with the probability scaling as $2s^2$. Finally, when neofunctional alleles are rare in the population, a functional allele can mutate to a neofunctional allele, and this can rescue a duplication that otherwise would be fated to be lost under drift alone. The first two paths thus fix already existing alleles, while the last requires the appearance of a new allele.

Linkage plays a key role in deciding which of the above selectively-driven paths are most likely. For a completely-linked duplication, only *ff* alleles can become fixed, as a completely linked *nn* duplication is lethal as a homozygote (the duplication process is assumed to be such that completely linked *nf* alleles cannot be generated). Thus, for complete linkage, neofunctional alleles play no role in the initial founding event, rather they must arise later by mutation. For large N, the simulation studies of Lynch et al., show that $\Gamma \simeq \Theta$, with both scaling

with population size. This occurs becomes essentially all fixed neofunctional alleles arise by mutations in duplications that otherwise would be lost from the population. The appearance of a neofunctional mutation rescues these duplications, driving them to fixation.

When the ancestral and duplicated loci are unlinked, selection no longer restricts successful initial duplications from involving neofunctional alleles. When neofunctional alleles are segregating at nontrivial frequencies in the population (requiring a sufficiently large population size and $\mu_c < s^2$), then neofunctionalization can occur by either a neofunctional allele founding, and fixing, the duplication or by fixation of a functional allele at the duplicate locus driving a neofunctional allele to fixation at the ancestral locus. Both of these probabilities occur with order $2s^2$, giving (for large N)

$$\Theta = \gamma \simeq 8Ns^2, \quad \text{and} \quad \Delta \simeq 4Ns^2. \tag{17}$$

If the population is still very large, but mutational pressure to nulls overwhelms neofunctional alleles at a single locus ($\mu_c > s^2$), then the main routine to neofunctionalization is through rescue of duplications that otherwise would go to fixation, but that acquire a neofunctional mutation that then drives them to fixation. Thus when the population size is large, but neofunctional selection is weak relative to null mutation, the probability of a map change is very small as the vast majority of neofunctional alleles are fixed at rescued new duplications, as opposed to the ancestral locus. Hence under these conditions neofunctionalized genes will almost always be at different locations than the original locus.

One final important result from the simulation studies is that fixation probabilities are much smaller under complete linkage compared to unlinked loci. Contrary to the two-fold increase in fixation probabilities under tight linkage for subfunctionalization, tight linkage results in orders-of-magnitude decreases in the probability of neofunctionalization compared to free recombination.

Implications for micro- and macroevolution

All three potential outcomes following a recent duplication (silencing, neofunctionalization, subfunctionalization) have importantly evolutionary consequences. Without silencing, the constant mutational pressure from gene duplication would continue to expand the genome. Further, cycles of duplication and silencing result in genomic dispersion of genes over time. Since this can modify the regulatory constraints on a gene (which can be very location-dependent), the dispersion of genomic location may reduce pleitropic constraints, promoting some degree of increased evolutionary flexibility. Subfunctionalization offers a potentially very significant chance to reduce the evolutionary constrains imposed of a multifunctional gene, and also results in genomic dispersion of function. Neofunctionalization is the engine of major adaptation.

All of these fates are interconnected, in that a round of duplication and gene silencing may reduce the evolutionary constraints on a gene, facilitating future neofunctionalization. Neofunctional alleles that cannot accommodate all of the original functions of the gene may exist in the population at low frequencies when a single locus is present, but can burst to fixation after a duplication appears, especially in large populations. It is quite likely that gene families have gone through cycles of subfunctionalization to disperse multiple functions from some original gene, followed by increased neofunctionalization given the reduced constraints in these new genes. Likewise, continual cycles of neofunctionalization can result in a collection of functions being accrued over time by a single gene. Duplication followed by subfunctionalization disperses these functions over several genes, offering further opportunities for additional evolutionary novelty to evolve.

Hence, it would appear that both subfunctionalization and neofunctionalization are critical, and complementary, events that shape the adaptive radiation of a gene family. What do the models tell use about the conditions under which these are favored? First, population size is critical. In small to medium populations, neofunctionalization is a rare event and whether silencing or subfunctionalization occurs is to a large degree a function of the gene complexity, either in terms of regulatory elements or dispersal of functions over distinct coding elements. More complexity (i.e., more independently mutable subfunctions) facilities subfunctionalization. Second, the relative strengths of mutation and selection are key elements, but these also interact with population size. In a small to medium population, if subfunctional null mutations are more frequency than silencing mutations, subfunctionalization is favored. However, as the population size becomes sufficiently large, a subfunctional mutation drifting towards fixation acquires additional mutations

that most likely silence it, driving the fate of genes strongly towards silencing. In small populations, neofunctionalization is highly unlikely unless both the selection coefficient and neofunctional mutation rate are significant. However, for a sufficiently large population, neofunctionalization is the dominant fate, either by fixing existing neofunctional alleles segregating at low frequencies or by rescuing a duplication that acquires a neofunctional mutation that then drives it to fixation.

Given that the optimal situation for the adaptive radiation of a gene family would be cycles of subfunctionalization to reduce evolutionary constraints followed by cycles of neofunctionalization starting with these less-constrained ancestors, a situation where a population experiences a long bottleneck to allow for subfunctionalization followed by a population expansion to facilitate neofunctionalization would seem to be optimal for macroevolution.

A final point that we will mention in passing is the suggestion that genomic dispersion of function can be a significant component towards reproductive isolation (Lynch & Force, 2000b). If cycles of duplication and silencing are occurring at a sufficiently high rate, then the genomic location of loci can change over relatively short periods of evolutionary time when populations are isolated. Ongoing genomics projects should shed light on the plausibility of this scenario.

Acknowledgments

Many thanks to Mike Lynch, Allan Force, and Jay Taylor for insightful discussions and to Manyuan Long and J.J. Emerson for very helpful comments on the manuscript.

References

Allenbdorf, F.W., 1979. Rapid loss of duplicate gene expression by natural selection. Heredity 43: 247–258.

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

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

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

Ferris, S.D. & G.S. Whitt, 1977. Loss of duplicate gene expression after polyplodisation. Nature 265: 258–260.

Ferris, S.D. & G.S. Whitt, 1979. Evolution of differential regulation of duplicate genes after polyplodization. J. Mol. Evol. 12: 267–317.

Ferris, S.D., S.L. Portnoy & G.S. Whitt, 1979. The roles of speciation and divergence time in the loss of duplicate gene expression. Theoret. Pop. Biol. 15: 114–139.

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

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y. Yan & J. Postlethwait, 1999. Preservation of duplicate gnees by complementary, degenerative mutations. Genetics 151: 1531–1545.

Haldane, J.B.S., 1933. The part played by recurrent mutation in evolution. Am. Nat. 67: 5–19.

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

Kimura, M. & 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.

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.

Lynch, M. & A. Force, 2000. The origin of interspecific genomic incompatibility via gene duplication. Am. Nat. 156: 590–605.

Lynch, M. & A. Force, 2000a. The probability of duplicate gene preservation by subfunctionalization. Genetics 154: 459–473.

Lynch, M., M. O'Hely, B. Walsh & A. Force, 2001. The probability of preservation of a newly arisen gene duplicate. Genetics (in press).

Maruyama, T. & N. Takahata, 1981. Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicate loci. Heredity 46: 49–57.

Nei, M., 1970. Accumulation of nonfunctional genes on sheltered chromosomes. Am. Nat. 104: 311–322.

Nei, M. & A.K. Roychoudhury, 1973. Probability of fixation of nonfunctional genes at duplicate loci. Am. Nat. 107: 362–371.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Heidelberg, Germany.

Ohta, T., 1980. Evolution and Variation in Multigene Families. Springer, New York.

Ohta, T., 1983. On the evolution of multigene families. Theoret. Pop. Biol. 23: 216–240.

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

Partridge, C.W.H. & N.H. Giles, 1963. Sedimentation behavior of adenylo-succinase formed by interallelic complementation in *Neurospora crass*. Nature 199: 304–305.

Spofford, J.B., 1969. Heterosis and the evolution of duplications. Am. Nat. 103: 407–432.

Takahata, N. & 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.

Walsh, J.B., 1985. How many processed pseudogenes are accumulated in a gene family? Genetics 110: 345–364.

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.

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